REPORT RESUMES

ED 002 910 24
THE RELATIONSHIPS BETWEEN SUCCESS IN GROUP COUNSELING AND
DISCREPANCY IN LEVELS OF PERSONALITY.
BY- CARLSON, WILLIAM A.
ILLINOIS UNIV., URBANA, COLL. OF EDUCATION
REPORT NUMBER CRP-623 PUB DATE 61
EDRS PRICE MF-$0.18 HC-$4.24 106P.

DESCRIPTORS- *GROUP COUNSELING, COUNSELING GOALS, *STUDENT IMPROVEMENT, ACADEMIC ACHIEVEMENT, ADJUSTMENT (TO ENVIRONMENT), *UNDERACHIEVERS, GIFTED, ADOLESCENTS, *PERSONALITY STUDIES, MENTAL HEALTH, EMOTIONAL MALADJUSTMENT, *INTERPERSONAL RELATIONSHIP, GRADE 9, URBANA, ILLINOIS

THE APPLICATION OF GROUP COUNSELING TO EDUCATION WAS INVESTIGATED WITH RESPECT TO ANALYSES OF INTERPERSONAL RELATIONSHIPS AND PERSONALITY INTEGRATION FOR USE WITH DEVELOPMENT AND MAINTENANCE OF SOUND MENTAL HEALTH PROGRAMS. SUBJECTS WERE 15 NINTH-GRADE STUDENTS, DIVIDED INTO 2 COUNSELING GROUPS AND CONSIDERED TO BE GIFTED UNDERACHIEVERS. EACH GROUP MET SEPARATELY FOR A TOTAL OF 16 SESSIONS. CRITERION MEASURES OF SUCCESS IN GROUP COUNSELING WERE OBTAINED FROM THREE BASIC RATING SCALES AND THE "PICTURE STORY TEST" (A FANTASY LEVEL PROJECTIVE EXAMINATION). THESE WERE TESTED FOR RELIABILITY AND THEN RELATED TO MEASURES OF DIFFERENCES BETWEEN THE SUBJECTS' LEVELS OF PERSONALITY. IT WAS FOUND AFTER THOROUGH STUDY THAT THE CRITERION OF SUCCESS IN GROUP COUNSELING WAS NOT SIGNIFICANTLY RELATED TO THE DISCREPANCY BETWEEN LEVELS OF PERSONALITY AT THE CONCLUSION OF COUNSELING. HOWEVER, THE USE OF SUCH A SMALL SAMPLE REQUIRED THAT THIS CONCLUSION BE REGARDED AS HIGHLY TENTATIVE. (JH)

ED002910

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

THE RELATIONSHIPS BETWEEN SUCCESS IN GROUP COUNSELING AND DISCREPANCY IN LEVELS OF PERSONALITY

A Supplement to
Cooperative Research Report No. 623

William Arthur Carlson

College of Education
University of Illinois
Urbana, Illinois
1961

The research reported herein was suggested through the Cooperative Research Program of the Office of Education, United States Department of Health, Education and Welfare. The report is based upon Mr. Carlson's doctoral dissertation in the Graduate College of University of Illinois. The Chairman of Mr. Carlson's committee was Merle M. Ohlsen. The other members of his committee were Thomas N. Ewing, William M. Gilbert, J. Thomas Hastings, Fred C. Proff, and William O. Stanley.

TABLE OF CONTENTS

Chapter Page

I REVIEW OF THE LITERATURE 1

  Historical Background 1
  Summary of Historical Background 6
  Empirical Background 7
  Summary of Empirical Background 15

II ANALYSIS OF SMITH'S DATA 18

  Summary 28
  Conclusions 29

III METHOD 30

  The Experimental Setting 30
  Subjects 30
  Levels of Personality 32
  Instruments 32
  Raters 36
  Rater Training 39
  Experimental Procedures 41
  Criterion 43
  Summary 44

IV RELIABILITY STUDY 46

  Validity and Reliability 46
  Analysis of "I" Responses 47
  Computing Reliability 48
  Rater Reliability 49
  Elimination of Items and Raters 54
  Criterion Reliability 55
  Item D Reliability 56
  Summary 60

V RESULTS AND DISCUSSION 62

  Procedural Summary 62
  Discrepancy Measures 62
  Criterion 63
  Testing Hypothesis I 66
  Testing Hypothesis II 72
  Summary 88

3414A
154

Chapter Page

VI SUMMARY AND CONCLUSIONS 90

Summary 90
Conclusions 92

REFERENCES 94

LIST OF TABLES

Table Page

1 Principal Axis Factors for the Seven Variables with a Varimax Rotation 20

2 Rotated Factors with the Loadings in the Same Order as the Correlation Matrix, i.e., by Dimensions of Meanings 22

3 Rotated Factors with the Loadings Ordered by Tests 24

4 Squares of the Factor Loadings Ordered by Dimensions of Meaning, Using the Principal Axes with a Varimax Rotation 25

5 Squares of the Factor Loadings Ordered by Tests, Using the Principal Axes with a Varimax Rotation 26

6 Behavior Inventory Rater Reliability 51

7 Need Behavior Scale Rater Reliability 52

8 Behavior Rating Scale Rater Reliability 53

9 Reliability of Criterion Rankings 55

10 Need Behavior Scale: Reliability of the Item D 57

11 Behavior Rating Scale: Reliability of the Item D 58

12 The Behavior Inventory: Reliability of the Item D 59

13 Intercorrelations Between the Averaged Criterion Rankings for All Fifteen Subjects 64

14 Intercorrelations Between the Averaged Criterion Rankings for the Thirteen Subjects 64

15 Agreement Among and Between Criterion Judges 67

16 Correlations Between the Averaged Criterion Rankings and the Discrepancy Between Levels I-A, I-B, and III on The Behavior Inventory 69

17 Correlations Between the Averaged Criterion Rankings and the Discrepancy Between Levels I-A, I-B, and III on The Need Behavior Scale 70

Table	Page
18 Correlations Between the Averaged Criterion Rankings and the Discrepancy Between Levels I-A, I-B, and III on The Behavior Rating Scale	71
19 Correlations Between the Criterion Rankings and the Discrepancy Between Levels I-A, I-B, III and the Self Ratings on The Behavior Inventory	73
20 Rank Order Correlations Between the Status Ratings of Adjustment and the _Pre_ _Minus_ _Post_ Differences of the D Values for The Behavior Inventory	75
21 Rank Order Correlations Between the Status Ratings of Adjustment and the _Pre_ _Minus_ _Post_ Differences of the D Values for The Need Behavior Scale	76
22 Rank Order Correlations Between the Status Ratings of Adjustment and the _Pre_ _Minus_ _Post_ Differences of the D Values for The Behavior Rating Scale	76
23 Rank Order Correlations Between the Criterion Rankings and the _Pre_ _Minus_ _Post_ Differences	78
24 Pre and Post Means for Item D Values and Pre Minus Post Mean Differences	80
25 Correlation Between the Item D Values Pre and the Item D Values Post	82
26 Rank Order of Growth and the Difference Between Pre and Post D Values for the Lay vs. P.S.T. Ratings on The Behavior Inventory	84
27 Rank Order of Growth and the Difference Between Pre and Post D Values for the Clinical vs. P.S.T. Ratings on The Behavior Inventory	85

CHAPTER I

REVIEW OF THE LITERATURE

Historical Background

Twenty-two years after Freud's death and sixty-one years after The Interpretation of Dreams, there is still disagreement among experts as to which of his contributions to science is the greatest (Jones, 1953, 397). There is little doubt, however, that his formulation of the concepts of the unconscious mind, and the role of conflict in emotional maladjustment has had a marked influence on the attempts to understand and treat mental disorders.

A short paper "Neurosis and Psychosis," which Freud published in 1924, sums up the relationship of conflict and psychopathology. He made it clear that he thought that both neuroses and psychoses "originate in the ego's conflicts with the various powers ruling it. . ." and that these conflicts may be either internal with id and/or super-ego, or external with the environment (outer world). Freud took the position that a neurosis represents the outcome of a conflict between the ego and the id; narcissistic neuroses, such as melancholia, one between the ego and the super-ego; the psychoses, one between the ego and the outer world (Freud, 1957, 250-254).

The theoretical formulations concerning maladjustment since Freud's initial attack upon the problem have been many and varied; however, the central issue has continued to be that of conflict, either within the individual's personality structure and/or with the outer world. Making this issue central has resulted in an automatic acceptance of a multi-level conception of personality. A short review of some of the major theorists will serve to demonstrate that they believe that conflict has been a major source of psychopathology, emotional disturbance, or personal maladjustment.

Adler has often been interpreted as maintaining a strict adherence to the concept of the unity of personality and repudiating the notion of intrapsychic conflict (Munroe, 1955, 432 and Way, 1950, 52, 240-245). However, it appears clear that even Adler's explanation of maladjustment actually describes man as caught in an intrapsychic conflict:

> The feeling of insignificance, of weakness, of anxiety and helplessness, of ill health, of deficiency, of pain, etc., produces in the neurotic actions of such a nature that he seems to be compelled to set up a defense against effeminacy, that is to say, to be obliged to act in a manly and forceful manner (Adler, 1917, 100-101).

This describes the "masculine protest" which is compensation for the feelings of inferiority which are equated with feminity. Later, he stressed that the feelings of inferiority stem from a sense of imperfection or feelings of insufficiency which result in a conflict-attitude, "the struggle for existence" or "struggle of all against all" (Adler, 1929, 13).

Jung described the total personality as being composed of many separate but interacting systems which, in the maladjusted person, are at war with each other:

> I want to stress that although the pathogenic conflict is a personal matter it is also a broadly human conflict manifesting itself in the individual, for disunity with oneself is the hall-mark of civilized man. The neurotic is only a special instance of the disunited man who ought to harmonize nature and culture within himself (Jung, 1953, 18).

Though Rank's earlier writings placed great importance upon "the birth trauma" (Rank, 1929), his later writings placed considerable emphasis upon the existence of a "will conflict" (Rank, 1950). When a person seeks therapeutic help, this conflict between wanting his own way and at the same time not wanting it becomes underscored:

> The help-seeking patient brings his whole neurotic conflict, which at bottom is a will conflict, to the analytic situation, to which he wishes to submit himself but which at the same time he resists. On this very

> conflict, the inability to submit and the inability to put over his own will positively, his whole neurosis depends (Rank, 1950, 16).

Karen Horney takes the position that, while it is not neurotic to have conflicts, the degree of the conflict, the degree of the subject's awareness or lack of awareness of the conflict, and the degree to which a person is able to make a choice between possible solutions mark the characteristic differences between normal and neurotic conflict (Horney, 1945, 31-33). The problem is not at all completely internal for the "neurotic is one who has experienced the culturally determined difficulties in an accentuated form. . . and who has consequently been unable to solve them. . ." (Horney, 1937, 290).

Some learning theorists have accepted this basic notion and placed considerable stress on the "social factors" which produce conflict.

> An intense emotional conflict is the necessary basis for neurotic behavior. The conflict must further be unconscious. As a usual thing, such conflicts are created only in childhood. . . neurotic conflicts are taught by parents and learned by children (Dollard and Miller, 1950, 127).

They then outlined in considerable detail the "social conditions for learning of unconscious conflicts" (Dollard and Miller, 1950, 127).

Mowrer was more explicit concerning the role of "social conflict." He first pointed out that not all individuals have the necessary skill, courage, and social opportunity to deal openly and realistically with their conflicts. At first they use deception and denial in relation to others, but under continued pressure of anxiety these techniques are used in relation to the self, i.e., repression.

> In psychotherapy, the object is to replace dissociations by associations, to substitute normal anxiety for neurotic anxiety, on the assumption that then, and only then, are the ordinary developmental processes able to operate and to bring the individual's conflicts to a constructive rather than pathological conclusion (Mowrer, 1950, 559-561).

Therapy results in the internal neurotic conflict being transformed into a social conflict so that the unfinished business of making the human animal into a normal, mature, morally and socially accountable human being can be completed

(Mowrer, 1950, 520-521).

Adler was the first to stress the importance of social or interpersonal factors in relation to mental disorders; Horney, Fromm, and Erik H. Erikson added further theoretical advancements (Hall and Lindzey, 1957, 115), (Leary, 1957, 10). Sullivan's posthumous publications are the most recent and in time may prove to be the most significant of these social-psychological theories. The impact of Sullivan's theories arises from his consolidation of a personality theory which is grounded in interpersonal relations.

As a result of early infantile experience with rewards, mild to moderate degrees of anxiety, and sudden severe anxiety there comes an initial personification of three phases of the self-system:

> . . . the beginning personifications of me are good-me, bad-me, and not-me. So far as I can see, in practically every instance of being trained for life, in this or another culture, it is rather inevitable that there should be this tripartite cleavage in personifications. . . (Sullivan, 1953, 161-163).

Good-me is related to the organization of experience in which satisfactions have been enhanced by rewards of tenderness or love. Bad-me is related to experience associated with increasing degrees of anxiety. Not-me comes from the experience of intense anxiety and is later encountered only in an occasional dream or in severe mental disorders such as schizophrenia (Sullivan, 1953, 161-163).

This basic internal division of personification of me occurs in relationship with others, but Sullivan also stated clearly that

> in avoiding or minimizing the anxiety that is inherent in the unceasing struggle to protect the self-system from the diffuse referential process that cannot be admitted into awareness, various specific processes, the dynamisms, come into play (Sullivan, 1956, 4).

More simply, one might say that an individual develops defense systems in relationship to the anxiety resulting from internalized conflicts.

This anxiety had its origin in infancy, and for the infant there is no difference between anxiety and fear. In the infant a fear-like state can be induced (1) by violent disturbance of his zones of contact with circumambient reality, and (2)

> by certain types of emotional disturbance within the mothering one. From the latter grows the whole exceedingly important structure of anxiety, and performances then can be understood only by reference to the conception of anxiety . . . . Anxiety, as a phenomenon of relatively adult life, can often be explained plausibly as anticipated unfavorable appraisal of ones current activity by someone whose opinion is significant (Sullivan, 1953, 9 and 113).

The problem is both intrapersonal and interpersonal, for anxiety results in a disruption of the communicative processes, and a large part of mental disorder is the result of, and is perpetuated by, inadequate communications (Sullivan, 1953, XII).

Leary has constructed a theoretical system combined with a research program in personality evaluation. He took the interpersonal dimension of personality as a starting point in developing his theory and in designing research. In a short but complex definition he stated that interpersonal behavior is "Behavior which is related overtly, consciously, ethically, or symbolically to another human being (real, collective, or imagined)" (Leary, 1957, 4). The notion of conflict between levels of personality, be it intra- or interpersonal, was taken as the moving force directly related to emotional maladjustment. Leary pointed out the age old problem of the perverse inconsistency of human behavior and the protean complexity of the human subject. All of these difficulties are the result of variability within and between human personalities; thus, the problem of conflict among levels of personality becomes the key in the understanding of human personality.

The first step in the unraveling of man's complex nature was provided by the theory of unconscious motivation. When Freud postulated a multi-level theory of personality, he made the first systematic explanation of conflicts, ambivalence, and inconsistency:

> The psychoanalytic theory of personality and of neurosis is defined in terms of interlevel conflict, i.e., variability. There are certain

> motives at one level and certain other motives at another. Their interplay spells out the organization of personality. . . . Variability is involved in almost every aspect of personality that we study. Conflict, ambivalance, defense mechanisms, growth, regression, change, improvement-in-therapy--all these phenomena have the basic factor in common--one unit of measure varies in relation to another. . . . The amount and kind of variability is a most significant variable of human behavior. It defines the type and intensity of conflict. It determines the tendency to change or to maintain a rigid adjustment (Leary, 1957, 242-243).

As a part of his research program Leary has developed an "operational definition of five levels of personality according to the source of data" (Leary, 1957, 81). Level I (Public Communication) concerns the interpersonal impact of the subject on others--his social stimulus value. Level II (Conscious Descriptions) is the subject's view of himself and the world. This is obtained from interviews, autobiographies, check lists, and questionnaires. Level III (Preconscious Symbolization) is the subject's autistic, projective fantasy productions. This level has two sublevels: III Hero and III Other. This study is concerned with Leary's first three levels. The next two levels involve special difficulties which raise questions concerning their usefulness (Leary, 1957, 84) and are only mentioned in passing: Level IV (Unexpressed Unconscious) and Level V (Ego Ideal).

Summary of Historical Background

Freud made the first systematic statement concerning the relationship between emotional maladjustment and interlevel personality conflicts, thus establishing a multilevel theory of personality. Jung not only found discord and opposition within the human personality but also viewed conflict as a basic fact of life. Some of Adler's followers have claimed that he repudiated the construct of intrapsychic conflict; yet it appears clear that he described man as experiencing both an internal struggle and an external conflict. Adler should be given credit as the first to suggest that all maladjustment had some relationship to interpersonal conflicts. From Jung and Adler until the present time there has been an increasing interest in the role of interpersonal and intrapersonal conflicts in maladjustment.

Dollard and Miller classified "social conditions for learning of unconscious conflicts." Mowrer and Rank both described the process of psychotherapy in terms of the internal conflict being acted out in the therapy relationship and thus being transformed into a social conflict or will conflict. Horney and Sullivan gave special attention to the role of anxiety in the creation and maintenance of maladjustment.

Building on social psychological theories of personality development and maladjustment, Leary constructed a theory and a system of personality research. He combined the basic ideas of "interpersonal behavior," "multilevels of personality," "interlevel conflict," and the classification of "levels of personality according to the source of the data." The operational result of Leary's work provided the method of making measurements of interlevel conflicts in terms of intra- and interpersonal variables. After careful consideration of the problems involved, he stated that

> Adjustment in terms of the over-all personality organization consists in flexible, balanced, appropriate, accurate interpersonal behavior. In terms of the subdivision of personality--the levels of public interaction, perception, and private symbolisms--it consists of appropriate, accurate, and balanced interpersonal behavior respectively (Leary, 1957, 31).

Empirical Background

The research which has a direct relationship to this study may be divided into two major groups: (1) research concerning the relationship of conflict and emotional maladjustment and (2) research concerning levels of personality. The present study may be considered as a synthesis of these areas of research.

Although the relationship between conflict and emotional maladjustment in human beings has often been discussed and described, experimental work with the lower animals has produced more dramatic and significant results. Such results were possible because researchers were able to run tightly controlled experiments and were free to induce a conflictual situation which produced results even though it seriously damaged the emotional structure of the animal.

In 1914, I. V. Pavlov accidentally, and then later purposefully, produced the first experimental animal neurosis (Pavlov, 1927, 289). This occurred when a dog had been conditioned to salivate to a strong electrical shock. The animal responded with an "alimentary reaction" to a stimulus which normally would evoke an inborn defense reaction. When generalization was attempted, a sudden pathological behavior change occurred. A second planned experiment produced identical results (Pavlov, 1927, 291).

This phenomenon has often been repeated in many experimental situations by a host of investigators. A complete review of these studies is neither necessary nor practical. All of the studies have as a common factor the presence of a stressful situation in which conflicts and frustrations are present. They also have as a common result the disruption of the normal behavior of the animals which the investigators routinely have labeled as neurotic-like behavior.

Liddell found disturbances similar to those in Pavlov's dogs during his work with pigs and goats. Gantt has developed abnormalities in dogs which simulate human psychopathology. Working with cats, Masserman observed "neurotic" behavior such as fear or phobic responses, chronic anxiety, compulsive behavior, fixations, and efforts to secure fondling by the researchers. Cook's studies using rats produced many unusual behaviors which were non-goal directed. On the basis of their behavior he classified his animals' disturbances into inhibitory and excitatory-type neuroses. Maier, and later Maier and Wapner together, were able to produce violent convulsions involving clonic activity of all legs. One animal, however, became extremely passive and catatonic like, holding whatever posture it was placed in by the experimenter. The above studies have been successful in "disrupting the normal behavior of a fair proportion of the animals tested, and in this sense they may be regarded as neurosis-producing situations" (Maier, 1949, 140-149). It appears from these findings that non-normal, maladjustive behavior occurs in lower animals (dogs, cats, rats, pigs and goats) when faced with conflictual situations.

The study of experimental neurosis has been extended to higher animals who are more closely related phylogenetically to man. Masserman, an experimental physiologist, as well as a trained psychoanalyst, produced an experimental neurosis in a laboratory-raised male vervet monkey who had been trained to respond to various signals in order to secure food. The experimental neurosis was precipitated after three separate sudden presentations of a toy snake just as he was about to receive his food pellet (Masserman, 1955, 336-337).

Hebb reported an unplanned experiment with four chimpanzees, two of which developed serious neurosis or psychosis. Although he was unable to explain the origin of the behavior disturbances, Hebb took the position that experience may have been a decisive factor,

> but, if so, it operated in some way that is quite mysterious in the light of our present knowledge of behavior. Freudian theory has the credit of recognizing the existence of a kind of learning that causes, apparently, no immediate emotional disturbance and yet may contribute to one much later (Hebb, 1949, 250).

One conflictual element common to these studies is that between a fear, _e.g._, an electric shock, air blast, etc., and the satisfaction of some basic physiological need, _e.g._, food. Another element, and possibly the more basic problem, is the conflict involved in making a successful discrimination in perception in order to respond with the correct adaptive behavior.

In all of the experiments

> the circumstances resulting in disorganization can be described as a conflict between _excitation_, the tendency to make a response, and _inhibition_, the tendency to withhold that response. Normal behavior requires a delicate balance between responding and inhibiting processes. Discriminative strain is, therefore, essentially the same as conflict. . . . When both excitation and inhibition, or approach and avoidance, are evoked inharmoniously at the same time, an animal finds itself helpless in an

> incomprehensible situation. The plight of a human being facing an intensely motivated conflict is not dissimilar (Shaffer and Shoben, 1956, 117).

Liddel believed that the major factor in experimental neurosis studies has been the traumatizing procedure which

> as a method of training, it might exert upon the nervous system effects essentially similar to the traumatic influences of certain cultural factors in human life. . . that the perplexing problems which face the animal in the conditioned reflex laboratory represent restraints and pressures situated at the apex of a pyramid. Progressive restriction of liberty and correlative increase in pressures (similar to those exerted by society on the human individual) extended from the wild state, through domestication, to the too refined training of the laboratory (Hunt, 1944, 391-393).

Shaffer and Shoben, reviewing essentially the same data, concluded that

> In Lewin's terms the restrained animal cannot 'leave the field' to avoid the brunt of the conflict by effective escaping or distracting responses. People, too, succumb to conflicts when they are hemmed in by actual social barriers imposed by others or by the boundaries they accept for themselves by internalizing the demands of their cultures (Shaffer and Shoben, 1956, 118).

These conceptual schemes easily fit into Leary's system of levels of personality and are in harmony with the hypothesis that the variability or conflict between these levels of personality is related to emotional maladjustment (Leary, 1957, 242-243). Thus, one would expect to find that similar experimental neurosis studies with human subjects demonstrate the same relationships and results as have been found in the studies using lower animals.

There are reasonable restrictions concerning the use of human subjects for experimental neurosis experiments as extreme as the animal studies. However, there have been less harmful and less dramatic approaches to this problem using human subjects. Russian experimenters, using children as subjects, have reported studies in which emotional excitement and a general disruption of conduct were produced. Work with American college students has produced similar results, but less intensely.

> The most severe conflict was produced by the double approach-avoidance situation. . . . The blocking of the motor response was ordinarily accompanied by emotional or substitute behavior such as giggling, swearing, or appealing to the experimenter (Shaffer and Shoben, 1956, 118-119).[1]

Hypnotic techniques have been used either to implant a temporary conflict as an experimental procedure or as a clinical tool in the therapeutic treatment of disturbed individuals. Hypnotic methods have also been useful in demonstrating the existence of an on-going personality conflict (Bremman, 1947, 76; Watkins, 1949, 354; LeCron, 1954, 390; Weitzenhoffer, 1957, 361).

What has been described as "a miniature psychotic storm" was produced in an apparently normal medical student as the result of an hypnotically induced conflict (Brickner and Kubie, 1936, 467-487). In another study, as a part of an informal research-teaching program, some examples of "the psychopathology of everyday life" have been demonstrated through the use of hypnotic induction. In each instance the independent variable was a conflict situation in which the subject experienced opposing motivations or feelings, e.g., "subject was instructed to feel that smoking was a bad habit, that he both loved and hated it. . . ." In each situation the subject manifested defensive, maladjusted behavior (Erickson, 1939, 338-353).

The affective power and potency of an "experimental neurosis" in human subjects were vividly demonstrated by two studies in which an experimentally induced conflict resulted in the abrupt, dramatic and apparently lasting resolution of a major neurotic problem (Erickson, 1935, 34-50; McDowell, 1948, 168-177). The experimental conflict produced so much emotional distress in one of the subjects that he was immediately re-hypnotized and the complex was "removed" (Erickson, 1935, 42). It appears that in each of these studies the experimenter was successful in restructuring the internalized neurotic conflict so that it could be symbolized and verbally resolved in terms of the experimental conflict. Although such success demonstrates a positive, healthy, final behavioral change, it also supports the basic notion that conflict between levels of personality and emotional maladjustment are closely correlated.

---

[1] This is a secondary source. The primary references in order of appearance are (Kraenogorski, 1925, 753-786; Razran, 1933; Freeman, 1940, 117-120; Hovland and Sears, 1938, 477-493; Sears and Hovland, 1941, 280-286).

A recent well-planned and carefully executed study of subception and repression also demonstrated the effect of the hypnotic induction of a pseudo-experience associated with hostility and a destructive impulse. The pseudo-experience became the source of severe anxiety and/or clinical symptoms. Some of the subjects solved the conflict by giving in to the destructive urge. Others could not reach such a solution and consequently continued to express anxiety and symptoms. The experimenter "actually verified tachycardia by counting rate and witnessed the eruption of a skin rash on the hands of two Ss. . ." (Reyher, 1958, 107-108).

Using thirty "normal" subjects, Keet made use of each subject's individual areas of conflict. He found conflictual areas as symptomatized by the subject's performance on the first sixty words of the Jung word association test. As the result of a learning task involving critical words representing the subject's conflictual areas, it was found that twenty-five of the thirty subjects missed repeating the critical words after learning a second set of neutral words. Application of an interpretative type tyerapy to the conflict area resulted in the subject's being able to recall spontaneously the critical word. When the "microneurosis" was resolved, the critical word was easily recalled.

> As soon as these [associated processes] came into play, the necessary associative links were formed and the effective experience which determined the failure to recall the word became conscious. This served to resolve the conflict. The result was the restoration of normal psychological functioning since the subject could now recall under circumstances similar to those in which he had forgotten the first word, another, also in functional relationship with the same affective experience. . . .
> This also served to demonstrate the close analogy between the situation

in the experiment and real breakdown of behavior as seen in the psychoneuroses (Keet, 1948, 54-55).

A problem in the replication of Keet's study has occurred. Other researchers have not been able to find a significant difference between forgetting the critical words and forgetting the neutral words. It is possible, however, that this problem is related to the sample. The other researchers report "We have some preliminary evidence which suggests that persons of differing personality make-up may react differently to the traumatic word in a learning task of Keet's type. . ." (Grummon and Butler, 1953, 597). It is possible that Keet's subjects were, by chance, more disturbed or maladjusted than he realized and that he was thus dealing with substantial personality conflicts.

Through the use of hypnotic techniques in a study of the dream in psychotherapy, data were collected from two on-going individual psychotherapy cases. By the use of associational and scaling procedures in the form of the Semantic Differential it was possible to make ratings periodically in both the hypnotic and waking states:

> It was predicted that therapy patients should experience a wide discrepancy between the conscious and unconscious meanings of persons centrally related to conflict areas. . . ,

and that

> successful treatment should effect a reduction in the difference between conscious and unconscious levels of meanings (Moss, 1953, 138).

The first prediction was upheld; there were extensive differences between the ratings of interpersonal concepts made in the hypnotic and waking states during the initial stage of psychotherapy. The second prediction was also substantiated; the successfully treated case showed a reduction in semantic distance between waking and hypnotic ratings, while the unsuccessful subject did not. Although these are encouraging results, they are limited in application because of serious methodological difficulties, e.g., using only two subjects and finding suitable statistical methods. However, it has been a useful study in pointing the way for further research (Moss, 1953, 64-68 and 137).

The above studies dealing with human subjects are suggestive of the existence of different levels within the human personality. There also is evidence suggestive that "as conflict is resolved and integration achieved, there is a coalescence of these originally discrepant levels of meaning" (Moss, 1953, 1). This type of suggestive research evidence and the theoretical notions concerning emotional maladjustment provide both theoretical and empirical evidence to justify further research concerning levels of personality.

Since 1949, the Kaiser Foundation, Psychology Research Project, Oakland, California, has been engaged in a program to develop a system of personality research which would be adequately complex and empirically objective. The Foundation has established operational definitions of the levels of personality in terms of the sources of the data contributing to them. Level I (Public Communication) is based on _others_ ratings of the subject; Level II (Conscious Descriptions) is based on the subject's _self-ratings_; and Level III (Preconscious Symbolization) is based on others ratings of the subject's _fantasy productions_ (Leary, 1957, 76-82). The Foundation has studied both normal and emotionally disturbed subjects in interpersonal situations, _e.g._, group therapy and discussion groups. Thus it had social situations which, though specialized were "more closely resembling real life situations than most which are utilized for obtaining of data for personality research purposes" (Freedman, _et al._, 1951, 145).

In one of its more recent studies the Kaiser Foundation presented a method for measuring changes in several levels of personality before and after brief psychotherapy. The Foundation group was able to demonstrate a significantly greater change in the therapy patients than in the control group. However, they did _not_ demonstrate improvement; they demonstrated only that some type of change had occurred (Leary and Harvey, 1956, 128-129).

From theoretical formulations based on clinical observation, as well as from Leary's approach to personality assessment, an unpublished doctoral dissertation preducted that "the extent of interlevel conflict or the amount of interlevel discrepancy in the meaning of self is positively and significantly related to the degree of emotional disturbance in persons" (Smith, 1958, 8). However, the assumption that various levels of personality were represented in the tests was rejected because there was an alpha coefficient of .85 for six of the twelve variables. This was interpreted as meaning that these six variables did not represent independent dimensions and were not drawn from different levels of personality (Smith, 1958, 40-41).

Further analysis of Smith's data will be presented below. The analysis indicates that the assumption based on the alpha coefficient is in error. As will be described later, a factor analysis of the intercorrelations of all twelve variables of meaning suggest that at least two levels of personality are represented in these data. This is taken as further evidence that there are, in fact, different levels of personality and that they can be defined and measured.

Summary of Empirical Background

Pavlov accidentally discovered that lower animals will react to an experimentally induced conflict with pathological behavior changes. A host of investigators since have demonstrated repeatedly that when animals are placed in stressful situations where conflicts and frustrations are present, there is a disruption of normal activity which the investigators routinely label as neurotic like behavior. Masserman and Hebb have produced or described similar results with the higher animals, e.g., monkeys and chimpanzees.

Shaffer and Shoben, as well as Liddel, have drawn analogies between the traumatizing procedures in the animal studies and the plight of a human being facing an intensely motivated conflict situation. Although there have been some limited experimental neurosis studies with human subjects, the usual approach has been through hypnotic and/or clinical techniques. Erickson, and later McDowell, used hypnotically induced experimental neurosis to resolve major neurotic problems. Reyher demonstrated that a hypnotically induced conflict can produce severe anxiety and/or clinical symptoms, e.g., a psychosomatic rash.

Keet presented evidence that a "microneurosis" could be resolved as soon as the subject was able to make the necessary associative links to resolve the related conflict. Through the use of hypnotic techniques, Moss confirmed his predictions that a therapy patient would show a wide discrepancy between conscious and unconscious meanings which were centrally related to conflict areas and that successful treatment would effect a reduction in these discrepancies. However, since only two cases were involved, these results cannot be generalized.

Since 1949, Leary and his associates have carried on a research program in the development of an interpersonal multilevel and operational approach to personality measurement. They have consistently defined the levels of personality in terms of the source of the data which contributed to them. Including both normal and emotionally disturbed subjects in their studies, they have centered their attention on the measurement of individuals in interpersonal situations such as therapy or discussion groups. The results of their studies present evidence suggestive of the existence and measureability of personality levels.

There are many theoretical statements which postulate a relationship between emotional maladjustment and interlevel personality conflicts. The implications common in counseling and clinical lore are that the reduction in these conflicts will result in improvement in mental health. Many research findings indicate that the induction of conflictual situations in both the lower animals and human subjects results in maladjustive behavior. Other studies suggest that for human subjects when conflict is resolved and integration is achieved there is a reduction in the discrepant levels of meaning or behavior.

Previous studies also indicate that levels of personality can be operationally defined and measured. Thus the following hypotheses may be deduced:

HYPOTHESES: I. The greater the degree of success in group counseling, the smaller the discrepancy between levels of personality at the conclusion of counseling.

II. The greater the degree of success in group counseling, the greater the reduction in the discrepancies between levels of personality from pre to post counseling.

CHAPTER II

ANALYSIS OF SMITH'S DATA

Smith favored the theoretical position that "the basic conflicts within a person are among attitudes or centrally organized meaning systems. . ." rather than between impulses at one level and overt behavior at another level (Smith, 1958, 5-6). Thus he viewed interlevel personality conflict in terms of a discordance among meanings at various levels, rather than in terms of different sources of interpersonal behavior, e.g., Leary's "Public Communication," "Conscious Description," and "Private Symbolization" (Leary, 1957, 77-80). He adopted Osgood's dimensions of meaning and operationally defined the levels of personality in terms of specific psychological tests (Smith, 1958, 6-8). He assumed that when listed in the following order, the Ewing Personal Rating Form, Gilbert Self-Interview, Osgood Semantic Differential, and the Thematic Apperception Tests range from most to least consciously determined content (Smith, 1958, 27-28.)

In his study, the client's self-perception was independently estimated from each of the four tests in terms of Osgood's dimensions of Evaluation, Potency, and Activity. Thus he obtained for each subject test scores which represented twelve variables of meanings. Of these twelve, Smith found that seven were reliable at or beyond the .01 level of significance.

Since the between-test correlations were generally as large as the reliability coefficients for three of the tests, Smith suggested that there was one substantial general factor represented in these scores. When he computed an alpha coefficient from the matrix of intercorrelations and found it to be .85, he concluded that

> one general self-evaluative factor accounted for 85 percent of the variance in the equally weighted composite of six of the seven reliable variables. The assumption that various levels of personality were represented in the tests was consequently rejected (Smith, 1958, 39-42, 52).

While it is true that the intercorrelations of the six variables seemed to be generally positive and substantial, the seventh variable (T.A.T.Activity) had low negative intercorrelations with each of the other six variables of meaning. Thus it appeared likely that it would require at least two factors to account for the intercorrelations of the seven variables. Using the matrix of intercorrelations of all seven variables with ones as communalities, a principal axis factor analysis was computed on Illiac, the University of Illinois Digital Computer. Illiac was instructed to test the latent root for each factor extracted and to stop factoring when the root became less than one. This procedure has been suggested as a "practical" approach to both the "Communality Problem" and the "Common Factor Space Problem" in factor analysis (Kaiser, 1960, 9).

Three factors were extracted; the first two had roots equal to, or greater than, one. Only the first two factors were rotated and interpreted.

> When unities are inserted in the diagonals of the correlation matrix, the variance of each test is equal to +1.00. A latent root measures the size of a principal axis factor, for a latent root is equal to the variance removed by the factor. Under these conditions, it does not seem reasonable to accept a factor which has less variance than a test. Thus, the number of factors which are nontrivial is also equal to the number of latent roots which are equal to or greater than +1.00 (Dickman, 1960, 66).

The first two factors after rotation by the Varimax Method (Kaiser, 1958, 187-200; Harman, 1960, 301-308) were group factors, the second of which was bipolar (see Table 1). Factor I was a large group factor which generally had heavy loadings on five of the seven variables. That the T.A.T. Activity variable was distinctly separate from the other six variables in terms of Factor I demonstrated that the Ewing and Osgood tests, along with the ratings from the Gilbert Self Inventory, were different from the ratings of fantasy material. The

Table 1*

Principal Axis Factors for the Seven Variables with a Varimax Rotation

Variable	Factor I	Factor II
Ewing Potency	.832	-.050
Gilbert Evaluation	.360	-.635
Gilbert Potency	.818	-.128
Osgood Evaluation	.546	-.675
Osgood Potency	.556	-.497
Osgood Activity	.728	-.417
T.A.T. Activity	.074	.852

***The sixty-six pages of Appendices have not been included in this report. However, they are available in the microfilm ($2.75) or Xerox ($8.20) copy of the complete dissertation. (Order No. 62-470). The Xerox is processed to 6 x 8½ inches. They may be ordered from University Microfilms, Inc., 313 North First Street, Ann Arbor, Michigan.**

difference suggests that conscious description of self (Leary's Level II) is different from private symbolization, i.e., fantasy material (Leary's Level III). The interpretation of the bipolar Factor II was more difficult; it might suggest the dimension of defensiveness with its opposite signs for Level II vs. Level III.

Tentatively, the two factors were interpreted as representing conscious description of self (Leary's Level II) and defensiveness or resistance, respectively. However, this interpretation, as well as the one following, may be questionable because only forty-nine subjects were used. In spite of this limitation, a further analysis was considered important in order to obtain at least some suggestions as to whether there would be support for "Levels of Personality" in Smith's data, when all twelve variables were studied.

The matrix of intercorrelations for all twelve variables of meaning were obtained and factored by the same procedure previously described. For this matrix five factors were extracted, and the first four had roots equal to, or greater than, one. These four were then rotated and psychological interpretations were made in terms of their meanings as related to the subjects, all of whom were judged to be predominantly moderately to severely neurotic (Smith, 1958, 25). Both an orthogonal and an oblique rotation were performed using the Varimax Method and the Kaiser and Dickman[1] oblique solution (Harman, 1960, 301-308, 326). The Varimax and the K.-D. oblique solution produced similar loadings; the latter had more extreme values and appeared to be a slightly better simple structure. When the restriction of orthagonality was removed the factors did not become extremely oblique or intercorrelated. That the largest intercorrelation was between Factors I and IV (+.389) suggested that these two Factors were somehow related. However, the remaining five intercorrelations ranged from -.136 to +.112 and suggested a nonsignificant relationship.

---

[1] Hereafter abbreviated as K.-D.

Table 2

Rotated Factors with the Loadings in the Same Order as the Correlation Matrix, i.e., by Dimensions of Meanings

Variable Dimensions of Meanings	Test	Varimax Rotation				Kaiser-Dickman Oblique Rotation (Primary Factor Pattern)			
		I	II	III	IV	I	II	III	IV
Evaluation	Ewing	.806	.076	.267	-.199	.931	.084	.237	-.357
	Gilbert	.671	-.100	-.174	.134	.655	-.047	-.176	.023
	Osgood	.847	-.083	-.213	.268	.809	-.008	-.212	.129
	T.A.T.	.262	.763	.016	.314	.265	.808	.002	.283
Potency	Ewing	.218	-.729	-.146	.363	.076	-.688	-.099	.354
	Gilbert	.560	-.570	-.098	.062	.526	-.538	-.083	-.033
	Osgood	.699	-.160	-.087	.314	.648	-.098	-.075	.209
	T.A.T.	.003	.148	.727	.150	.047	.117	.739	.176
Activity	Ewing	.057	-.026	.052	.884	-.144	.043	.110	.943
	Gilbert	.209	-.601	.196	.246	.136	-.589	.233	.236
	Osgood	.437	-.383	-.227	.669	.256	-.298	-.180	.638
	T.A.T.	-.237	-.152	.803	-.223	-.137	-.232	.807	-.176

Primary Factor Correlations

	I	II	III	IV
I	1.0000	-.1362	-.1021	.3890
II		1.0000	.1118	-.1149
III			1.0000	-.1266
IV				1.0000

The variance which a factor has in common with a test is an indication of their relationship. This variance--found by squaring the factor loadings--is easily interpreted because the larger values maintain their relative size while the smaller ones become insignificant (Dickman, 1960, 104; Thurstone, 1947, 73). The general configuration is also more apparent when the factor loadings are arranged by tests rather than by the dimensions of meaning. Therefore, Tables 3 and 5 were used for the psychological understanding and interpretation of Smith's data.

From Tables 3 and 5, it is apparent that Factor I is a group factor with major loadings from one or more dimensions of meaning on the Ewing, Gilbert, and Osgood tests. With the Varimax rotation it has loadings of .81, .67 and .85 for the dimension of Evaluation; .56 and .70 for Potency; and .44 for the Osgood Activity. Thus it appears to be a factor associated with the Evaluation, Potency and Osgood Activity dimensions. In addition, it is clearly separate from the T.A.T., which has only trivial loadings, and thus suggests a more basic psychological interpretation of Factor I. The Ewing, Gilbert and Osgood were all direct measures of the clients' statements or ratings concerning themselves, while the T.A.T. was fantasy productions. It thus appears that the most basic dimensions being measured by Factor I are those of conscious self description, _i_._e_., Leary's Level II (Leary, 1957, 78-79).

The Primary Factor Correlations from the K.-D. oblique rotation listed in Table 3 showed that five of the intercorrelations were trivial and that there was only a slight departure from orthogonality. This was _not_ true for the sixth correlation between Factors I and IV, which was +.3890 and indicated obliqueness (arc. cos. .3890 = $67^{o}6'$). This similarity between Factors I and IV was also shown by their insignificant T.A.T. loadings as well as by their heavy loadings on some of the dimensions of meaning from the Ewing and Osgood tests. That Factor IV had loadings of .88 on Ewing Activity and .67 on Osgood Activity

Table 3

Rotated Factors with the Loadings Ordered by Tests

Variable Test	Dimensions of Meaning	Varimax Rotation I	II	III	IV	Kaiser-Dickman Oblique Rotation (Primary Factor Pattern) I	II	III	IV
Ewing	Evaluation	.806	.076	.267	-.199	.931	.084	.237	-.357
	Potency	.218	-.729	-.146	.363	.076	-.688	-.099	.354
	Activity	.057	-.026	.052	.884	-.144	.043	.110	.943
Gilbert	Evaluation	.671	-.100	-.174	.138	.655	-.047	-.176	.023
	Potency	.560	-.570	-.098	.062	.526	-.538	-.083	-.033
	Activity	.209	-.601	.196	.246	.136	-.589	.233	.236
Osgood	Evaluation	.847	-.083	-.213	.268	.809	-.008	-.212	.129
	Potency	.699	-.160	-.087	.314	.648	-.098	-.075	.209
	Activity	.437	-.383	-.227	.669	.256	-.298	-.180	.638
T.A.T.	Evaluation	.262	.763	.016	.314	.265	.808	.002	.283
	Potency	.003	.148	.727	.150	.047	.117	.739	.176
	Activity	-.237	-.152	.803	-.223	-.137	-.232	.807	-.176

Primary Factor Correlations

	I	II	III	IV
I	1.0000	-.1362	-.1021	.3890
II		1.0000	.1118	-.1149
III			1.0000	-.1266
IV				1.0000

Table 4

Squares of the Factor Loadings Ordered by Dimensions of Meaning, Using the Principal Axes with a Varimax Rotation

Variable		Factor I	Factor II	Factor III	Factor IV	Communalities	Per Cent of Variance	
							Per Testing Unit	Per Dimension of Meaning
Evaluation	Ewing	.65	.01	.07	.04	.77	9.3	34.5
	Gilbert	.45	.01	.03	.02	.51	6.1	
	Osgood	.72	.01	.04	.07	.84	10.1	
	T.A.T.	.07	.58	.00	.10	.75	9.0	
Potency	Ewing	.05	.53	.02	.13	.73	8.8	30.8
	Gilbert	.31	.32	.01	.00	.64	7.7	
	Osgood	.49	.02	.01	.10	.62	7.4	
	T.A.T.	.00	.02	.53	.02	.57	6.9	
Activity	Ewing	.00	.00	.00	.78	.78	9.4	34.8
	Gilbert	.04	.36	.04	.06	.50	6.0	
	Osgood	.19	.15	.05	.45	.84	10.1	
	T.A.T.	.06	.02	.64	.05	.77	9.3	
	Factor Variance	3.03	2.03	1.44	1.82	8.32		
	Per Cent of Total Variance	36.4	24.4	17.3	21.9		100.1%*	
					100.0%			

---

*Rounding error.

Table 5

Squares of the Factor Loadings Ordered by Tests, Using the Principal Axes with a Varimax Rotation

Variable		Factor I	Factor II	Factor III	Factor IV	Communalities	Per Cent of Variance	
							Per Testing Unit	Per Test
	Evaluation	.65	.01	.07	.04	.77	9.3	
Ewing	Potency	.05	.53	.02	.13	.73	8.8	27.5
	Activity	.00	.00	.00	.78	.78	9.4	
	Evaluation	.45	.01	.03	.02	.51	6.1	
Gilbert	Potency	.31	.32	.01	.00	.64	7.7	19.8
	Activity	.04	.36	.04	.06	.50	6.0	
	Evaluation	.72	.01	.04	.07	.84	10.1	
Osgood	Potency	.49	.02	.01	.10	.62	7.4	27.6
	Activity	.19	.15	.05	.45	.84	10.1	
	Evaluation	.07	.58	.00	.10	.75	9.0	
T.A.T.	Potency	.00	.02	.53	.02	.57	6.9	25.2
	Activity	.06	.02	.64	.05	.77	9.3	
	Factor Variance	3.03	2.03	1.44	1.82	8.32		
	Per Cent of Total Variance	36.4	24.4	17.3	21.9		100.1%*	
					100.0%			

*Rounding error.

marked it as measuring an Activity dimension. Thus, Factor IV, like Factor I, appears to measure the more basic psychological dimension of conscious self description, i.e., Leary's Level II.

That Factor III is a T.A.T. Potency-Activity measure was shown by respective loadings of .73 and .80 on those variables. All other loadings were immaterial as shown by the squares of the loadings in Table 5. This establishes Factor III as a measure of strength and action expressed at the fantasy level of self description. It is also highly consistent with the category, "Private Symbolization," which is described as being made up of projective, indirect, fantasy materials and has been labeled as a Level III measure of personality (Leary, 1957, 79).

Because of its bipolarity, Factor II required a more complex interpretation. Its largest loading on the Varimax rotation was T.A.T. Evaluation +.76; the remaining important loadings had a negative sign. The Ewing and Gilbert Potency variables were -.73 and -.57; the Gilbert and Osgood Activity loadings were -.60 and -.38. Again the Level II and III dimensions appeared, but this time in terms of the sign changes, or the bipolarity, of the factor. In psychological terms it appears that, when the fantasy productions (T.A.T. stories) of a group of moderately to severely neurotic clients were judged as being good, they rated themselves as weak and inactive (Ewing and Osgood) and that their verbal statements about themselves (Gilbert) were rated as showing weakness and inactivity.

When the signs for the loadings on Factor II were reflected, the only difference was a reversal in the descriptive terms.[2] When the moderately to severely neurotic clients' conscious self descriptions (Level II) indicated strength and action, their fantasy productions (Level III) indicated badness.[3] Leaving the signs for the loadings as they were in the Varimax rotation gave the following results: when the clients' conscious self descriptions (Level II) indicated weakness and lack of action, their fantasy productions (Level III) indicated goodness. It should also be noted that the dimensions of T.A.T. Evaluation and T.A.T. Potency and Activity are not related (Compare Factors II and III).

---

[2] The reflection of the signs is purely arbitrary and does not change the adequacy of the solution (Harmon, 1960, 110, 173-174).

When the loadings on Factor II were taken as they originally occurred, the factor could be considered as a measure of "blandness." In the clients' terms, when they described themselves as weak and inactive, their fantasy productions were rated as good. With the signs reflected, the factor could be considered as a measure of "aggressiveness." When the clients described themselves as strong and active, their fantasy material was rated as bad. One psychological concept which would include both of these possibilities is defensiveness. It also would be theoretically correct to expect this mechanism to operate across both levels (Leary's II and III) of personality. This then would be a measure of neurotic defensiveness.

Summary

Analysis of Smith's data failed to find one general factor, but rather has found four factors; the largest factor only accounted for 36.4% of the variance in the factor matrix (Table 5). If one used Leary's operational definitions of levels of personality, the Ewing, Gilbert, and Osgood tests would be regarded as measures of Level II; the T.A.T., as a measure of Level III. Upon this theoretical basis, the analysis does demonstrate the existence of Level II

---

[3] Evaluation was defined for both the Gilbert and T.A.T. ratings as Good vs. Bad. Good as: more good, favorable, attractive, pleasant, valuable, clean, fair, nice, honest, or happy; Bad as: to appear worse, unfavorable, unattractive, unpleasant, worthless, dishonest, dirty, cruel, unfair, sad, distasteful (Smith, 1958, 61).

(represented by Factors I and IV) and Level III (represented by Factor III). These two levels are also represented in the bipolar Factor II, which is considered to be a measure of neurotic defensiveness.

Conclusions

Analysis of Smith's data failed to find one general factor.

It did find four factors which could be interpreted in terms of Leary's definitions of levels of personality.

Factors I and IV were interpreted as being a measure of personality Level II, or Conscious Self Description.

Factor III was interpreted as being a measure of personality Level III, or Private Symbolization.

Factor II, a bipolar factor, was interpreted as representing aspects of both personality Level II and Level III and was named Neurotic Defensiveness.

This analysis suggests that future attempts to measure or use "levels of personality" should follow Leary's operational definitions.

CHAPTER III

METHOD

The Experimental Setting

This study has been a part of a much larger long range project investigating group counseling with various age groups in educational settings (Ohlsen and Proff, 1960, 4-5). Educators have become increasingly aware of the need for developing and maintaining sound mental health and the importance of interpersonal relationships and personality integration. Thus there has been a move towards the application of group counseling in educational institutions (Broedel, 1958, 1-2).

A nationally prominent secondary school cooperated with a team of researchers from the College of Education at the University of Illinois in providing a setting where group counseling could be intensely investigated. All of the sessions of two counseling groups were kinescoped and these, along with the subjects' Picture Story Test protocols and self ratings on a behavior inventory, became the basic raw data for the present study. The use of the kinescope records has made it possible to study and rate the interpersonal behavior of the subjects under controlled conditions.

Subjects

The two counseling groups[1] consisted of fifteen ninth-grade students who were considered to be gifted underachievers in need of counseling. The first group was made up of two girls and six boys, while the second was composed of two girls and five boys. These groups met twice a week for eight weeks for a total of sixteen sessions. The same therapist worked with both groups. All of the subjects[2] were in the top ten per cent of their class as tested on the California Test of Mental Maturity and were at least one or more deciles lower in their grade point averages (Noble, 1958, 26-27; Ohlsen and Proff, 1960, 8-9).

---

[1] No attempt has been made to differentiate between the theory and practice of counseling and psychotherapy. However, the term counseling is used to indicate that the clients were treated in an educational, rather than a medical setting.

From the beginning of the Project, there has been complete agreement among the research team members that underachievement is not a surface phenomenon and can not be easily changed. There has also been general agreement that, though the subjects might be considered normal adolescents, they were experiencing many difficulties which are typical of underachievers and approach neurotic-like behavior. Most of the clients questioned whether they were intellectually gifted, apparently feeling that the pressure for academic success (school grades) was so intense that they had to deny their ability. They also often appeared hostile, unsociable, overly self-sufficient, indifferent to responsibilities, negative towards school work, and generally hard to reach (Ohlsen and Proff, 1960, 25-26).

Previous research findings, based on an analysis of the data by groups, indicated that the first group had made significant growth while the second one had not made such growth. Two detailed studies of the interpersonal behavior of the clients in the second group revealed that it contained two subjects who were extremely hostile and interfered with the progress of the group. However, there were also two different clients who contributed most to making the counseling a therapeutic experience (Ohlsen and Proff, 1960, 23-41, 69-95).

---

[2] Throughout this paper, the term client and subject are used interchangeably.

Levels of Personality

It is difficult, if not impossible, to prove the existence of levels of personality in terms of the traditional conscious vs. unconscious dichotomy or typically Freudian conceptualizations. A simpler approach is to define the levels of personality in terms of the sources of the data (Leary, 1957, 75-83). Thus Level I (Public Communication) is the overt interpersonal behavior of the subject as rated by others. Level II (Conscious Descriptions) is the subject's view of himself and others. Level III (Private Symbolization) is the subject's production of projective, indirect fantasy materials. This method is appealing because it is both parsimonious and practical to expect that the ratings of a subject's public interpersonal behavior would be somewhat different from his own self descriptions, and that both of these would be a great deal different from ratings of his private or extremely personal expressions. The analysis of Smith's data as presented in Chapter II lends credence to this approach by demonstrating factor differences between Levels II and III.

The major part of this study has been concerned with the differences between the ratings of the subjects in group counseling (Level I) and the ratings of the subjects' fantasy productions based on a Thematic Apperception Test[3] type of projective test (Level III). However, one of the three rating scales used did provide a measure of conscious self description, or Level II; it had been a part of the original project research and the subjects rated themselves on it.

Instruments

This study used four instruments: three rating scales[4] and the Picture Story Test,[5] a T.A.T. type of fantasy level projective test. Five pictures made up the P.S.T.: Card II of Murray's T.A.T., Card I of the Michigan Picture Test, and Cards I, IV, and V of Alexander's and Cronbach's Adaptation of the T.A.T.[6]

---

[3] The Thematic Apperception Test will hereafter be abbreviated T.A.T.

[4] Copies of the rating scales used in this study are in Appendix B.

[5] The Picture Story Test will be abbreviated P.S.T.

The original research team believed that these five pictures were especially good in depicting "situations which might be expected to be within the experience of the students and which might be expected to elicit stories involving interpersonal relationship with parents, school personnel, and peers as well as self attitudes" (Broedel, 1958, 43). In brief, the P.S.T. was relevant to the types of interpersonal difficulties which the subjects were experiencing. Although the P.S.T. was administered in a group setting, the procedures followed allowed the subjects to finish their stories and thus produced usable results (Broedel, 1958, 43-44). This writer later administered the P.S.T., using the same procedures, and found that all subjects finished their stories within the time limits.

The Behavior Inventory[7] was developed as a part of the original project and as a measure of interpersonal relationship. It consisted of twelve behavior categories which were paired in every possible combination to give a total of sixty-six items. Such a form resulted in a forced-choice instrument which was expected to reduce response set and/or "halo effect" (Broedel, 1958, 47). Almost all, if not all, the raters who have used the B.I. have disliked it because they were forced to make choices between paired favorable-unfavorable as well as between paired favorable and between paired unfavorable items.

For use with Leary's theoretical approach to levels of personality, it was ideal that the B.I. was designed to measure interpersonal behavior and based on revisions of Bales' Interaction Analysis (Bales, 1950; Broedel, 1958, 45-46). Leary and his associates found that their interpersonal variables and Bales' variables had considerable overlap even though they had been developed independently.

---

[6] Positive Prints of these pictures are in Appendix C.

[7] The Behavior Inventory will be abbreviated B.I.

> It is interesting to note this. . . especially since Bales as a sociologist may perhaps be described as more interested in social interaction as such than in the individual personality structure of the participants or subjects, the latter being the main focus of attention of the research herein described (Freedman, et al., 1951, 149).

During the initial phases of the research project the subjects had rated themselves on the B.I., thus providing this study with one measure of conscious description or Level II.

Two forms of the B.I. were used: one was used as a self description by the clients or subjects; and the other, as a rating instrument by the kinescope and P.S.T. raters. The two forms were identical, except for the form of the pronouns. The first was cast in first person, present tense; the second, third person, present tense.

The Behavior Rating Scale[8] used in this research is the product of several revisions. The first was made by Broedel and Ohlsen, who replaced the forced choice format with a seven point Likert type scale. The twelve basic descriptive statements were increased to forty-one descriptions of behavior or characteristics (Gazda, 1959, 73-77). The new form resulted in a considerable improvement and was useful as a measure of mental health (Ohlsen, Proff, and Southard, 1959, 12-13). However, the seven point rating scale required the rater to make judgments which were too exact, and thus resulted in a lower test retest reliability (Gazda, 1959, 76).

A second revision was undertaken by the research staff when they reduced the rating scale from seven to five points and added thirty-two items (making a total of seventy-three items). Maslow's theories of motivation and personality (Maslow, 1954) and Jahoda's report on criteria of positive mental health (Jahoda, 1958), served as the basic guide lines in the development of the new items. The project director had made extensive notes in his study of Maslow's theories and Jahoda's report which were then translated into the descriptions of behavior and personality characteristics. A careful inspection of the B.R. will demonstrate that the majority of the items are descriptions of behavior and the ones which might be classed as personality characteristics are only one inference step away from a behavior description.

---

[8] The Behavior Rating Scale will be abbreviated by B.R.

The third revision was made by this writer in collaboration with the project director. The B.R. was rewritten in the third person, present tense, so that it could be easily used by the rating teams. The major change, however, was in the wording of the directions to the raters. The second revision of the "self" form had five points on the rating scale which required the rater to compare the individual, on a percentage basis, to people in general of his age group. The third revision contained a statement indicating simply how accurate or inaccurate the item was in describing the individual. This was done in an attempt to focus the rater's attention directly on the subject's behavior in terms of each item. An additional "I" response was also added which indicated "insufficient information to make a judgment."

The Need Behavior Scale[9] was a thirty-five item rating scale based on the conceptual framework developed by Stern and his associates in their work on personality assessment. Their conceptual framework was based mainly on the work of Murray, who defined a list of needs for use in T.A.T. analysis (Stern, et al., 1956, 69-73; Stein, 1955, 50-58). The items for the N.B. scale were originally constructed from these two lists of needs and their definitions. These items were then independently judged by three of the project research staff members for clearness and rateability. Five items were dropped because they were judged to be too idiosyncratic. Several of the remaining thirty-five items were modified slightly during the training of the clinical psychologists as kinescope raters. The final form of the N.B. scale was then used by all three of the rating groups.

---

[9] The Need Behavior Scale will be abbreviated N.B.

Raters

Fourteen raters were used in this study: nine highly trained clinical psychologists and five psychologically unsophisticated laymen. All of these raters were paid at a rate which was commensurate with their regular employment. Such payment was a marked departure from the usual procedure of either "talking fellow professionals into exchanging help" or making what is basically a token payment. Thus, the basic motivation common to the total group was money. This extrinsic goal was needed because the ratings were long and difficult; most raters worked from thirty-five to forty hours to complete the tasks. It seems certain that few, if any, of the raters could have "been talked into" giving so much of their time and energy without a reasonable payment for their services. The nine professional clinical psychologists did demonstrate considerable interest and concern, above and beyond any monetary rewards, that they would be reliable and accurate in their judgments. The raters were divided into three groups; five clinical psychologists (group one) and five secretaries (group two) rated the subjects from the kinescope recordings and four clinical psychologists (group three) rated the subjects P.S.T. stories.

The five clinical psychologists who rated the kinescopes[10] ranged in age from the late twenties to the late thirties. There were four men and one woman; all except the female rater received his professional academic training at the University of Illinois. The female rater had taken her academic training at another large university, but had been in the same clinical psychology training program as the other four psychologists in a Veterans Administration Neuropsychiatric Hospital.[11] Two of the five clinical raters had completed their Ph.D. degrees in Clinical Psychology and the other three were in the advanced phases of their academic work. One of the raters was a Clinical Psychologist in a County Mental Health Clinic, and had been a staff member at the V.A.N.P. Hospital for a short time after completing his training at that station.

---

[10] Hereafter abbreviated as clinical raters.

The four P.S.T. raters were all male clinical psychologists ranging in age from the late twenties to the late thirties. Three of them did their academic work at the University of Illinois and were in the Clinical Psychology training program at the V.A.N.P. Hospital. The fourth received his academic training in another large university and had been in the Counseling Psychology training program in a different V.A.N.P. Hospital. However, he was a staff member as a Clinical and Counseling Psychologist at the same V.A.N.P. Hospital where all of the other professional raters had received their clinical internship. He had completed his Ph.D. degree, and the other three P.S.T. raters were in the advanced phases of their academic work. One of the P.S.T. raters also was employed part time as a Clinical Psychologist in a County Mental Health Clinic.

Since there were five clinical raters, it would have been good procedure to have had five P.S.T. raters; however, the supply of such high level professionally skilled personnel simply was not sufficient. It also would have been good procedure to have had both groups rate the kinescopes and the P.S.T.

---

[11] Hereafter abbreviated as V.A.N.P. Hospital.

However, the great amount of time involved in completing either set of ratings made this impossible. Indeed, it was considered fortunate that it was possible to obtain the services of these highly trained professional people for the amount of time and effort required for the procedures used. As the above descriptions indicate, the two professional groups - the five clinical raters and four P.S.T. raters - were considered equally competent clinical psychologists. Thus it was assumed that, had the two groups been interchanged, they would have produced equivalent ratings of the subjects in the kinescope recordings or the subjects' P.S.T. stories.

The five psychologically unsophisticated laymen[12] who rated the subjects from the kinescopes were female secretaries who ranged in age from the early twenties to the middle thirties. The two criteria used in selecting them as raters were that they had no formal training in psychology and were willing to do the work. Four of the five lay raters had not attended college, let alone taken any work in clinical psychology. One had about three semesters of college work and had taken one course in introductory psychology. Since the course was a very general survey of the whole field of psychology and did not place special emphasis on clinical, counseling, or related fields of psychology, it was decided that she could still be considered a psychologically unsophisticated layman for the purpose of making these ratings.

All of the lay raters were sure they would not be able to rate the subjects because of their lack of training. When they were told that they were being asked to do the ratings because of their lack of training, they became willing to undertake the work, especially in view of the monetary reward. The lay raters group differed from the professional rater groups in that they were all females, psychologically unsophisticated, and more motivated by the monetary aspects than by the desire to be competent professionally. What experimental differences, if any, resulted from the sex difference is not known. However, other researchers have found that, apparently regardless of sex differences, nonprofessionals provide even more reliable ratings of human behavior than professionally trained persons. Leary and his coworkers have found that naive, psychologically untrained raters have surpassed psychologists and psychiatrists in making reliable Level I ratings (Leary, 1957, 112-115). As the ratings of the kinescopes proceeded, it appeared that the lay raters had become interested in the task and were also greatly involved in "doing the best job possible."

---

[12] Hereafter abbreviated as lay raters.

Rater Training

Training raters on practice material until all had a common frame of reference, or perceptual set, was a critical step in obtaining reliable and, therefore, meaningful ratings. The general procedure was to provide the raters detailed definitions and descriptions of the behavior they were asked to rate. The instructions[13] for the B.I. were presented in written form, separate from the items, while the N.B. scale and the B.R. scale contained this information as part of the items. Then the directions and items were discussed until the raters said that they understood the material. Next they rated practice material, compared ratings, discussed reasons for differences, and agreed on how they would rate similar material in the future. These procedures were repeated until the raters attained significant agreement.

The training material for the clinical raters were four sound motion picture films from the film library of the University of Illinois Audio-Visual Aids Service.[14] The first film served as material for making initial ratings on the B.I. and B.R. scales, as well as for making final minor changes in the

---

[13] See Appendix D.

[14] Descriptions of these films are in Appendix E.

N.B. scale items. The three remaining films were the basic training films. An analysis of the ratings of the first film indicated that the clinical raters were attempting to make ratings which were several clinical inferences away from the observed behavior. It appeared that they were rating some of the items in the way they thought the subjects would rate them (Level II): others as they thought the "subject really feels," and the remaining items as if they were data from the subjects' fantasy materials (Level III). This problem has been reported by other researchers (Leary, 1957, 113-114). Corrective action was taken in this study through the training procedures, by having the clinical raters limit themselves to the observed behavior. This appears to be a growing practice, for similar instructions were given to the psychologist-psychiatrists rating teams in the current Veteran's Administration Chemotherapy Project No. 6 (Lorr, et al., 1960, 2).

The inclusion of the lay raters was a direct attempt to obtain a set of ratings which would be as unilevel as possible in the belief that the psychologically unsophisticated person tends to rate only what he sees and hears. In brief, they do not "psychologize" (Leary, 1957, 113). Training procedures for the lay raters were identical to those for the clinical raters except that only the three basic training films were used. These three films pictured different types of problems in interpersonal situations among high school age students, and thus were somewhat similar to the type of kinescope material which the raters later viewed for the research. The lay raters were required to rate various students from these films and their ratings were compared and discussed.

The P.S.T. raters were supplied with a random selection of sets of stories written by members of two earlier counseling groups. The P.S.T. raters were told only the grade (ninth) and sex of the writers. This procedure was also followed for the experimental stories. The P.S.T. raters followed the same training procedures as the clinical and lay raters; they rated a set of stories from a subject, compared the ratings, discussed the differences, and agreed how similar material would be rated in the future.

More freedom to make psychological inferences was allowed the P.S.T. raters because they dealt with projective material and had to develop a psychological picture of the type of individual who would create such a set of stories. They rated five stories as a group and thus had to develop a "generalized hero" for all five stories. It was this "hero" which was then rated. This approach, while different from most story by story ratings of the T.A.T. type of fantasy material, is actually the kind of organization which a clinical psychologist has to accomplish in order to integrate any blind analysis of projective material. Typically, the T.A.T. analysis results in the same type of organized picture of the client, patient, or subject. Thus the P.S.T. raters were asked to make such an integration before rating each set of stories. The results of the training session clearly indicated that such a procedure was possible and would produce reliable results.

Experimental Procedures

The kinescope films[15] made at the beginning and end of the group counseling were viewed jointly by the clinical and lay raters. They then, individually and independently, rated the subjects before the next viewing session, until all four sets, pre and post for both groups, had been viewed and rated. Three forty-minute kinescope films from one of the counseling groups made up one set for each viewing session. These sets of films were presented in random order.[16] During the viewing sessions the raters made notes of the subject's behavior which they later used in making their ratings. They were supplied with enlarged positive prints made from one of the 16mm kinescope movie film frames for each of the two groups. This made identification of the subjects as certain as possible. The raters were also given a random order to follow in their rating of subjects and were told that the four viewing sessions were presented in random order and that the subjects were in a random order.

---

[15] See Appendix F for the exact film session numbers.

[16] See Appendix G for the random order used.

The raters were asked to rate each subject on all three instruments in the following order: B.I., N.B., and B.R. This order was followed because the raters found the B.I. extremely difficult and fatiguing as is usual with forced-choice rating scales. It appeared that this order would reduce the fatigue effect and increase the possibility of obtaining reliable ratings. The comments made by the raters during the training sessions concerning their dislike of and difficulty with the B.I. also indicated that following the above order was a wise procedure. However, following this procedure introduced the possibility of a "halo effect" occurring (Thorndike, 1920, 25-29; Johnson and Vidulich, 1956, 130-134).

Two major corrective actions, directly related to the ratings, were taken to reduce or prevent the occurrence of the "halo effect." First, the forced-choice technique used in the B.I. helped to reduce the "halo effect" (Lindzey, 1954, 362-363). The major "halo effect" thus might take place between the N.B. and B.R. scales. However, the training period had stressed judging the subject on each item separately from the other items because such a procedure would help to negate the "halo effect" (Jahoda, 1951, 207). During the data analysis and hypothesis testing, all of the comparisons were made on one instrument at a time. No comparisons or tests have been made between instruments. Therefore, even if a "halo effect" from instrument to instrument happened to occur, it did not become a factor in the data analysis.

All sets of five P.S.T. stories per subject, both pre and post counseling, were randomly ordered before the P.S.T. raters evaluated and rated them.

At this point in the rating procedures it was discovered that by some unexplainable error, the sets of stories for two of the subjects were not in the project files. An intensive search failed to find them and, as of this date, they have not been located. It was decided that the amount of work which had already gone into the clinical and lay ratings justified continuing the study. The project research staff also believed that the loss of these two subjects would not be crucial. The number of subjects for the P.S.T. ratings was reduced to thirteen. The twenty-six sets of stories gave the raters the impression that they were rating a total of twenty-six subjects. The only direct information given to the raters concerning the subjects was their grade (ninth) and sex.

It is good clinical practice when rating or evaluating responses to projective materials to match the responses to the original stimuli material. This often adds meaning to the response which otherwise would be missed. Therefore, the P.S.T. raters were supplied with five by seven positive prints of each of the five slides used with the subjects. These prints had also been used in the training of the raters.[17]

The P.S.T. raters evaluated and rated the sets of stories individually and independently. They also followed the routine described above in rating each set of subject's stories on the B.I., M.B., and B.R. scales. The rationale for this rating procedure has already been discussed above.

Criterion

In order to determine the relationship between the success in group counseling and the discrepancy in levels of personality, it is necessary to have some independent measure of success. For the purpose of this study, the definition of the amount of success in group counseling was described in terms of the degree to which a client "made the most mental health growth, improvement, positive behavior changes, etc." Using this definition, a number of ratings were obtained by having the raters rank the clients from the extremes to the middle - from the ones who "grew most and least" to the middle until all of the clients were ranked.[18]

---

[17] See Appendix C.

Individual, independent rankings were obtained from the Group Therapist and Project Clinical Observer, as well as from the clinical raters, lay raters, and the P.S.T. raters. An additional check on these ratings was obtained by using the results of a previous status rating on the mental health adjustment of the subjects.

It is clear that the criterion ratings of the Group Therapist and Project Clinical Observer and the results of the separate mental health ratings would be independent of the ratings in this study. It is equally clear that criterion rankings by the clinical, lay and P.S.T. raters are only independent of the discrepancy scores which were obtained from personality level measures which did not involve their own kinescope or P.S.T. ratings.

Summary

As a part of a larger, on-going research project concerning group counseling, two sets of kinescope recordings selected from the beginning and ending sessions of two counseling groups were viewed and the subjects independently rated by five clinical psychologists and five secretaries. The pre and post sets of stories written to five T.A.T. type of pictures were studied and independently rated by four clinical psychologists. These ratings were made on three rating scales: the B.I., developed from a Bales' type of description of behavior; the N.B., a rating of "needs" as originally defined by Murray and revised and used by Stein and Stern; the B.R., an extensive revision of the B.I., with many of its items based on Maslow's psychological theories and Jahoda's review of the criteria of mental health.

---

[18] See Appendix H for the detailed instructions.

Operational definitions of levels of personality based on the sources of the data (Leary's approach) were used. The levels were defined as follows: Level I is others ratings of the subject's "public behavior," Level II, self ratings; and Level III, the ratings of the subject's fantasy productions. This approach has formed the basis of the data analysis; comparisons and tests were made between the levels of personality, but always on the same rating instrument.

Criterion measures of success in group counseling were obtained from several different sources and were tested for reliability and then related to measures of differences between the subject's levels of personality. This last operation formed the basis of the hypothesis testing.

CHAPTER IV

RELIABILITY STUDY

Validity and Reliability

Determining the validity of ratings is a difficult and troublesome problem. When ratings are used as the criterion measure, they can not actually demonstrate their own validity. However, ratings are often used because they are the only instrument available for obtaining the desired measure, even though such use evades the problem of validity. Current expert comments are of little help in solving this dilemma: "the validity of the rating procedure should be established in the particular situation where it is used" (Cronbach, 1960, 518). An earlier suggestion solves the dilemma by implying that no one has found a better answer: "Observations by impartial observers are generally accepted as valid if they can be made reliable" (Cronbach, 1949, 401).

There are some logical guide lines which one might follow in making estimates of the maximum possible validity with an independent criterion when the number of cases (the N) is extremely large. Considering that the validity cannot exceed the square root of the reliability (Cronbach, 1960, 131-132) and that the square of a correlation indicates the variance common to both score distributions (Diamond, 1959, 195-197), then, $r_V = \sqrt{r_R}$, and $r_V^2 = r_R$.[1] Since r^2 is equal to the proportion of the variance of each variable which is directly associated with the variance of the other variable (Lindquist, et al., 1951, 675), then under the best possible conditions the reliability correlation coefficient indicates the maximum amount (upper bounds) of variance which could be accounted for if an independent criterion were known. This statistical relationship adds logical support to the notion that reliable ratings may be assumed to be valid. Although the relationship indicates neither what the criterion would be, nor in what way the ratings would be valid, it does suggest what the maximum possible validity would be when the N is extremely large.

---

[1] r_V = validity correlation coefficient and r_R = reliability correlation coefficient, with the N extremely large.

The relationship between reliability and validity clearly indicates the importance of obtaining reliable ratings. Since the reliability between judges, taken individually, are often discouragingly low, it has been considered a good procedure to combine the ratings of several judges. "In the average the bias of one judge tends to cancel the bias of another, and each adds information the other had no opportunity to observe" (Cronbach, 1960, 510-511). The procedure of combining raters has been tested empirically and found to be successful:

> The results of this experiment tend to confirm this claim. The ratings of individuals when factored, it was shown, would not have been entirely factorially valid. This is perhaps due primarily to the loss of reliability rather than to the individual's inability to perceive the fundamental traits embedded in a set of variables. For when the ratings were combined into groups of three, the factors obtained were shown to be factorially valid (Dickman, 1960, 115-116).

The reliability of the ratings was computed between the individual judges or raters and were then averaged to obtain an estimation of the mean correlation for each group of raters. Reliability coefficients were also computed across subjects for the item D values[2] when the differences were between the ratings of individual judges on each item and also when the differences were between sets of combined (means of the judges' ratings) ratings on each item.

Analysis of "1" Responses

The directions given the raters for marking the items on the N.B. and

---

[2] See Chapter V for a description of the item D procedure.

B.R. scales included an "I" response by which the raters indicated that "there was insufficient information to make a judgment. The client did not demonstrate behavior which would permit one to rate him on this item."[3] The purpose of this response was to prevent the raters from marking the mid-point on the rating scale when, in fact, they were indicating a "don't know" response. "I" responses were scored as zero, *i.e.*, *not* rated, and the total of "I" responses was subtracted from the total of all responses. Since the per cent of regular scale ratings, per rater, ranged from 86 per cent to 100 per cent, with the mean being 96.6 per cent,[4] it is clear that the raters generally had sufficient information on which to base their ratings.

Computing Reliability

The ratings generated a mass of data which appeared to be beyond analysis. The B.I. totaled 26,664 responses for all raters, subjects, and items. The N.B. and B.R. scales added another 43,632 responses for a grand total of 70,296 responses. There were also fourteen rankings of growth and seven status rankings of emotional maladjustment. The number of the data and the inclusion of the statistically difficult "I" response made necessary the use of high speed computers for most of the data analysis.

The "I" response resulted in zero scores occurring in the regular rating scale scores on a random basis throughout the data. This problem was solved by the use of the *Incomplete Product-Moment Correlation Program* on the I.B.M. 650 Digital Computer. This program instructed the computer to include the paired X and Y values in the computation of the correlation *only* when both X and Y values were greater than zero. Using this program, the product-moment correlations[5] between raters were computed both across items and across subjects, *i.e.*, r^s with items as N and r^s with subjects as N. This procedure resulted in the computation of a grand total of 3,300 r^s for the pre and post reliability of the raters. Interpretation in meaningful terms of so many correlations was not possible. A computer program[6] was written especially for this problem to obtain average correlation coefficients.

---

[3] See Appendix B.

[4] See Appendix I.

This program first converted the correlations to Z' values by using a "Table of Z' Values for r" (Edwards, 1950, 409) and then averaged the Z' values by taking a weighted average of the Z'^{s}. The Z'^{s} were weighted by the inverse of their sampling variance (McNemar, 1955, 148; Edwards, 1950, 133-134). The averaged Z' value was then converted back into an r by again reading from the "Table of Z' Values for r."[7] This reduced the 3,300 r^s to 990 r^s, 165 of which were the most basic for the judgments made concerning the reliability of the raters.

Rater Reliability

The r between each pair of raters was computed across items on subjects and across subjects on items for both the pre and post ratings on the B.I., N.B., and B.R. rating instruments. These r^s were then averaged by means of the "r to Z'" computer program described above. Mean correlations were thus obtained for each rater's intercorrelations with all other raters on each subject and on each item. A mean correlation was also obtained for each subject and each item. When an $\bar{r}$[8] is reported for a subject or an item, it indicates a grand average for all[9] correlations for each group of raters on that subject or item.

---

[5] Hereafter the product-moment correlation will be abbreviated by r.

[6] This I.B.M. 650 program was written by Mr. Bud J. Meador, Assistant Director of the University of Illinois Statistical Service Unit, and is available for general use.

[7] Even greater accuracy and, with some computers, greater speed could be obtained if the Z' value were obtained by having the computer find Z' directly using: $Z' = \frac{1}{2}\,[\mathrm{Log}_e\,(1 + r) = \mathrm{Log}_e\,(1 - r)]$ (Edwards, 1950, 126; Peters and Van Voorhis, 1940, 155). Expressed in logarithms to the base 10, then $Z' = 1.1513\ \mathrm{Log}_{10}\frac{1 + r}{1 - r}$ (McNemar, 1955, 147).

Tables 6, 7, and 8 list a summary on each rating instrument of the average correlations, both across items on subjects and across subjects on items. The "range on subjects" indicates the smallest and largest $\bar{r}$, both on the pre and post ratings. The "Mean," pre and post, is simply the "grand average" of all the raters: Lay, Clinical, *or* P.S.T., for all subjects or all items. The tables then are a general summary of the reliability as measured between raters on each rating scale, pre and post counseling.[10]

An inspection of those three tables indicated that there was a moderate interrelationship among the raters. While it would have been "statistically pleasing" to apply significance testing to the correlations, this was not done because of technical difficulties. The significance testing would have involved the estimation of the "degrees of freedom" or sample size from which the original correlation was computed. The sample size varied as a result of the application of the incomplete product-moment correlation routine. Thus, one might properly demand that each of the several thousand correlations be tested individually. A more reasonable approach would have been to test the significance of the mean correlations (McNemar, 1955, 148-149). However, this could not be correctly carried out without inspecting the several hundred correlations which were a part of the final average correlations to determine the "degrees of

---

[8] $\bar{r}$ indicates an average or mean product-moment correlation.

[9] That is, all raters in each rating group, *e.g.*, all Lay raters, all Clinical raters, or all P.S.T. raters.

[10] See Appendix J for the $\bar{r}$ for subjects and items.

Table 6
Behavior Inventory
Rater Reliability

	Range and Mean Correlations				Maximum Pairs on Which Any One r Could be Based, i.e., Maximum N
	Across Items on Subjects				
Raters	Range on Subjects*		Mean**		
	Pre	Post	Pre	Post	
Lay Raters	.30 .67	.26 .64	.367	.443	12
Clinical Raters	.29 .78	.32 .76	.497	.494	12
P.S.T. Raters	.18 .90	.22 .72	.485	.549	12
	Across Subjects on Items				
Raters	Range on Items*		Mean**		
	Pre	Post	Pre	Post	
Lay Raters	.20 .51	.21 .37	.330	.289	15
Clinical Raters	.28 .62	.25 .51	.403	.361	15
P.S.T. Raters	.16 .56	.28 .65	.334	.421	13

*Average of the Raters on One Subject or Item.
**Average of the Raters on All Subjects or Items.

Table 7

Need Behavior Scale
Rater Reliability

Range and Mean Correlations

Across Items on Subjects

Raters	Range on Subjects*				Mean**		Maximum Pairs on Which Any One r Could be Based, i.e., Maximum N
	Pre		Post		Pre	Post	
Lay Raters	.21	.52	.12	.32	.300	.230	35
Clinical Raters	.21	.65	.20	.56	.464	.444	35
P.S.T. Raters	.25	.59	.24	.72	.408	.463	35

Across Subjects on Items

Raters	Range on Items*				Mean**		
	Pre		Post		Pre	Post	
Lay Raters	.01	.53	.14	.94	.297	.324	15
Clinical Raters	.17	.77	.12	.64	.438	.428	15
P.S.T. Raters	.19	.64	.04	.82	.399	.454	13

*Average of the Raters on One Subject or Item.
**Average of the Raters on All Subjects or Items.

Table 8

Behavior Rating Scale
Rater Reliability

Range and Mean Correlations

Across Items on Subjects

Raters	Range on Subjects*				Mean**		Maximum Pairs on Which Any One r Could be Based, i.e., Maximum N
	Pre		Post		Pre	Post	
Lay Raters	.18	.62	.24	.50	.328	.344	73
Clinical Raters	.23	.62	.22	.57	.429	.420	73
P.S.T. Raters	.24	.62	.24	.60	.425	.387	73

Across Subjects on Items

Raters	Range on Items*				Mean**		
	Pre		Post		Pre	Post	
Lay Raters	.00	.80	.10	.95	.314	.346	15
Clinical Raters	.10	.75	.00	.91	.374	.408	15
P.S.T. Raters	.15	.84	.14	.90	.409	.420	13

*Average of the Raters on One Subject or Item.
**Average of the Raters on All Subjects or Items.

freedom." In addition, it is very probable that the majority, if not all, of the mean correlations would have proved to have been significant, had they been tested, because of the substantial increase in the "degrees of freedom."

Elimination of Items and Raters

As Tables 6, 7, and 8 suggest, there were a number of items on each of the rating instruments which had averaged correlations which were near zero. Thus a procedure was developed which eliminated approximately half of the items from each of the rating instruments. A frequency distribution was constructed for each rating scale based on the mean correlations for both pre and post ratings for all three rating groups. The 40th centile was then computed and an item was dropped if more than two of its six mean correlations were below the 40th centile value. Since all of these computations were carried out using Z' values, it was possible to compute new "final mean" correlations for the groups selected. This method resulted in the selection of 58 per cent (seven items) on the B.I., 54 per cent (nineteen items) on the N.B., and 52 per cent (thirty-eight items) on the B.R.[11]

A similar procedure was used in the dropping of raters. The mean Z' value was computed on subjects for each rater, and raters were dropped when their own mean Z' value fell below the final group mean Z' values. These computations were run on the new set of average correlations which had been computed across items on subjects after the items had been dropped by the item selection described above. In general, two raters were dropped from each of the "five man" rating groups and one rater was dropped from the "four man" rating group. The mean reliability of each rater determined the raters to be dropped. This elimination procedure made it possible to have three completely separate "sets" of hypothesis testing: first, with all raters and all items used, then, with all raters but with some items dropped, and lastly, with both some raters and some items dropped.

---

[11] See Appendix K.

Criterion Reliability

A different approach was used to determine the reliability or agreement, among the raters on their rankings of the client's "success in group counseling." With five Clinical and Lay raters, ranking fifteen subjects and four P.S.T. raters ranking thirteen subjects, Kendall's Coefficient of Concordance (W) was computed (Edwards, 1954, 404-407). The coefficient of concordance indicates the degree of agreement among the judges and is, therefore, an estimate of judges' reliability. The significance of the coefficient of concordance for each set of rankings was tested by the application of the Chi square test, X_r^2 (Edwards, 1954, 411-412).

Table 9

Reliability of Criterion Rankings

Judges	Coefficient of Concordance	Chi Square Test of the Significance of W	
Lay Raters	W = .6671	$X_r^2 = 46.70$	
		P = .000255	d.f. = 14
Clinical Raters	W = .583	$X_r^2 = 40.80$	
		P = .000255	d.f. = 14
P.S.T. Raters	W = .6592	$X_r^2 = 31.154$	
		P = .002792	d.f. = 12

All three groups of judges demonstrated sufficient agreement within their own group to make it possible to combine their separate rankings into sets of average ranks for each group of judges. These three sets of combined ranks, along with three other criterion measures were then used in the hypothesis testing.

Item D Reliability

Not only was it necessary to establish that the raters were generally reliable, but also that the item D values between the different levels of personality were reliable. Because it would have been an involved and time consuming task to test the reliability of the more than 400 item D^s between individual raters, it was decided to make selected tests. Therefore, the reliability coefficients were computed for the item D values between the three clinical raters and the three P.S.T. raters who had been judged as most reliable on the original raw scores. These correlations were computed on the reduced number of items, on each rating scale, which had also been found to be most reliable. Thus the odds favored obtaining item D values which were reliable. The item D values, as well as the reliability coefficients, were thus computed between the individual raters as well as between the mean values of the five raters vs. the mean values of the three raters. (See Tables 5, 6, and 7).

The individual item D values computed from the more reliable raters and items demonstrated low to moderate reliability and were judged large enough, for item D reliabilities, to be usable. When the mean values of the five and three man rating groups were used to compute item D values and were tested for their reliability, all of the correlations were above +.75.[12] However, there was confounding present in these data because the five man group contained the three man group. Nevertheless, the excessive amount of time needed to compute the mean values for the items, the item D values, and the correlations required some such practical approach.

---

[12] These correlations were obtained as part of the I.B.M. 650 Computer output in obtaining the correlations between all of the group item D^s and the Criterion scores.

Table 10

Need Behavior Scale: Reliability of the Item D

(N = 13)

Between Individual Raters

Pre Counseling	C_3P_1*	C_4P_3	C_5P_4	Post Counseling	C_3P_1	C_4P_3	C_5P_4
C_3P_1	-	.706	.125	C_3P_1	-	.584	.360
C_4P_3		-	.401	C_4P_3		-	.654

*C_3P_1 Indicates that the item D was obtained from the ratings of Clinical Rater 3 (C_3) and the P.S.T. Rater 1 (P_1).

Between Groups of Raters

Between All Five Raters and Three of the Raters

Pre Counseling	Post Counseling
.939	.821

*Item D computed between the mean value of the Clinical Raters and the mean value of the P.S.T.

For N = 13 and a One Tail Test

r	P
.476	.05
.553	.025
.643	.01
.684	.005

Table 11

Behavior Rating Scale: Reliability of the Item D
(N = 13)

Between Individual Raters

Pre Counseling				Post Counseling			
	C_2P_1	C_4P_4	C_5P_5		C_2P_1	C_4P_4	C_5P_5
C_2P_1	-	.166	.441	C_2P_1	-	.268	.317
C_4P_4		-	.717	C_4P_4		-	.658

Between Groups of Raters

Between All Five Raters and Three of the Raters

Pre Counseling	Post Counseling
.924	.775

Table 12

The Behavior Inventory: Reliability of the Item D
(N = 13)

Between Individual Raters							
Pre Counseling				Post Counseling			
	C_1P_1	C_2P_4	C_3P_4		C_1P_1	C_2P_2	C_3P_4
C_1P_1	-	.489	.024	C_1P_1	-	.142	.254
C_2P_2		-	.352	C_2P_2		-	-.006

Between Groups of Raters

Between All Five Raters and Three of the Raters	
Pre Counseling	Post Counseling
.843	.825

Summary

A statistical relationship between reliability and validity gives logical support to the notion that reliable ratings may be assumed to be valid. The technique of combining ratings of several judges in order to increase reliability, a procedure the validity of which has been confirmed by a recent experiment, has been used throughout this study.

An "I" or insufficient information response, was included on two of the three rating scales to prevent the marking of the midpoint of the scale when the rater actually was rating the item as a "don't know" response. A count of the "I" responses indicated that there generally had been sufficient information for the ratings to be made. Because of the mass of data to be analyzed and the randomly distributed "missing ratings" resulting from the "I" response, which were scored as zero, high speed data processing machines were used throughout most of the reliability study.

Rater reliability computed across items on subjects and across subjects on items for both the pre and post ratings demonstrated a moderate, but sufficient agreement among the raters. In an attempt to "cross validate statistically," unreliable items and raters were dropped. This elimination, which cut about half of the items from each rating instrument and about one-third of the raters, was carried out in two steps: first, items were dropped, then raters. Thus three completely separate tests of the hypotheses could be made: All raters and all items; then all raters but some items dropped, and finally with both some raters and some items dropped. The reliability of the criterion rankings was determined on a group basis and all three groups demonstrated significant agreement.

A low to moderate interrelationship was found for the item D values between selected individual raters. This relationship was judged to be sufficient for item D reliabilities, which usually are low in terms of their absolute values because they are a reliability of a difference score. A more substantial correlation was found between the five and three man rating groups, but confounding of the data prevented the interpretation of these correlations.

CHAPTER V

RESULTS AND DISCUSSION

Procedural Summary

The major premise of this study has been that there is an inverse relationship between success in group counseling and the discrepancy between levels of personality: that the greater the success in group counseling, the smaller the discrepancies at the completion of counseling and the greater the reduction in these discrepancies from pre to post counseling. The discrepancies between levels of personality were measured by the item D procedure and the product-moment correlation. These discrepancy scores or values were then related to the criterion measures of success.

Discrepancy Measures

The item D was used as the major measure of discrepancy between the levels of personality. Item D is basically a difference measure between items taken one pair at a time. D is the statistical measure of profile similarity or profile discrepancy (Fiedler, et al., 1958, 14-15; Osgood, 1957, 90-91). D^2 is first computed; its square root becomes the item D.[1] The product-moment correlation was also computed and interpreted as a second measure of discrepancy with +1.00 indicating no "discrepancy" and -1.00 indicating maximum "discrepancy."

---

[1] $D^2 = \Sigma X^2 + \Sigma Y^2 - 2 \Sigma XY$ (X is one set of item scores and Y is a second set of scores on the same items, i.e., an X variable and a Y variable).

The item D^{s}[2] and r^{s} (later converted to Z' values)[3] were computed between Levels I (kinescope ratings), II (self ratings), and III (P.S.T. ratings). Following the suggestions from previous research, the same variable system was used in measuring the different levels of personality (Leary, 1957, 43; Smith, 1958, 29). Thus all discrepancy measures were computed between Levels I-A, I-B, II, and III,[4] and, always, between the ratings completed on the same instrument or rating scale.[5]

Criterion

Measurement theory predicts, and empirical studies have demonstrated, that the reliability of ratings is improved when judges are combined (Cronbach, 1960, 510-511; Dickman, 1960, 115-116). Table 9 indicates that there is considerable agreement within the three groups who made criterion rankings of growth. Thus, the rankings within each group were combined. The average of the ranks for each group should then be the "best estimate" of the "true" rankings of success in the counseling (Siegel, 1956, 237-238). The intercorrelations[6] of these averaged ranks (a "best estimate") with three other criteria rankings are presented in Tables 13 and 14. The two largest correlations in each of the two tables are spurious because they represent the relationship between two sets of variables and a third variable which is the average of the first two. That

---

[2] The terms item D, D value, and D are used interchangeably.

[3] The r^{s} between the Levels of personality were converted to Z' values (Edwards, 1954, 305-307 and 503) and a +3 was added to each subject's Z' values for ease in using the I.B.M. 650 computer.

[4] I-A = Lay raters, I-B = Clinical raters, II = Self ratings, III = P.S.T. ratings.

[5] The B.I., N.B., and B.R.

[6] Product moment correlations.

Table 13

Intercorrelations Between the Averaged Criterion Rankings for All Fifteen Subjects

	Clinical Raters	Lay Raters	P.S.T. Raters	Average of T&CD	Clinical Director	Therapist
Clinical Raters		.679	.114	.092	.277	-.121
Lay Raters			-.008	.300	.609	-.114
P.S.T. Raters				.127	.119	.071
Average of Therapist and Clinical Director					.798	.807
Clinical Director						.289
For 13 d.f. and a one-tailed test:	P =	.05	.025	.01	.005	
	r =	.441	.514	.592	.641	

Table 14

Intercorrelations Between the Averaged Criterion Rankings for the Thirteen Subjects

	Clinical Raters	Lay Raters	P.S.T. Raters	Average of T&CD	Clinical Director	Therapist
Clinical Raters		.710	.190	.139	.352	-.107
Lay Raters			-.203	.245	.592	-.164
P.S.T. Raters				-.169	-.259	-.023
Average of Therapist and Clinical Director					.752	.804
Clinical Director						.214
For 11 d.f. and a one-tailed test:	P =	.05	.025	.01	.005	
	r =	.476	.553	.634	.684	

is, the "Average of Therapist and Clinical Director" variable is spurious when correlated with either the Therapist or the Clinical Director rankings.

There are two significant findings represented in these two tables. First, there is a significant positive relationship between the criterion rankings of the Lay and Clinical raters. This was expected because both of those rankings were basically a Level I measure and should, therefore, demonstrate considerable agreement. The intercorrelations of the P.S.T. criterion rankings could only be interpreted when the N = 13; therefore, only the Table 14 intercorrelations were meaningful for the P.S.T. relationships.[7] Of the five intercorrelations for the P.S.T. variable in Table 14, four are near zero and/or negative. Thus the relationship between the P.S.T. criterion and the other variables is generally near zero and/or in the negative direction. This result is similar to the findings concerning Smith's correlations discussed in Chapter II. Five of the variables in Table 14 were generated by Level I data ("public"), while the sixth variable had Level III (T.A.T. fantasy material) as a source. Again, as happened in Smith's data, the "fantasy ratings" appeared to be different from data which were more "public" in nature.

Growth ratings (ranked from "grew most" to "grew least") were used as the criterion measure, since the hypotheses were based on the assumption that success in group counseling can be described in terms of the amount of improvement or the degree of growth in mental health. However, using the growth ratings as the criterion measure failed to test whether the more maladjusted clients "grew most." Thus a separate set of criterion ratings which had originally been designed as a status measure of maladjustment[8] was used to test whether the client who was ranked as "grew most" was also the most maladjusted.

---

[7] There were P.S.T. stories for only 13 subjects. See Chapter III for the details concerning this problem.

[8] See Appendix L.

If this were so, one would expect to find a negative correlation between the two variables because the status ratings had been ranked to indicate adjustment. It was also desirable to obtain an estimate as to whether the different criterion groups had applied similar standards in making their judgments. Both of these questions could be tested by the application of Kendall's "coefficient of concordance" (Siegel, 1956, 237; Edwards, 1954, 412). A coefficient of concordance (W), an average rank correlation coefficient ($\bar{r}'$), and separate rank order correlations (r') were computed between the Lay and Clinical rankings of growth and the status ratings of maladjustment. The status ratings were ranked so that "one" indicated least maladjustment and corresponded with the growth rankings of "one" which indicated the most growth.

The results, as shown in Table 15, suggested that, in general, the raters had ranked the more adjusted clients as having grown most. These results also suggested that the Lay rankings, Clinical rankings and the status rankings of adjustment were applying similar standards in making their judgments.

Testing Hypothesis I

The first hypothesis predicted that the greater the degree of success in group counseling, the smaller the discrepancy between levels of personality at the conclusion of counseling. This was tested by correlating the criterion values with the discrepancy measures at the conclusion of the counseling. If this hypothesis were tenable, the criterion measures should be positively and significantly related to the D values and negatively and significantly related to the Z' values.[9] Specifically, it was expected that there would be a significant relationship between the criterion rankings and the D and Z' values obtained between personality Levels I-A and III and between I-B and III.

---

[9] The product moment correlations between the levels of personality were converted to Z' values and a +3 was added to each Z' value for each subject for ease in using the computer.

Table 15

Agreement Among and Between Criterion Judges

Relationship Between Status Ratings of Adjustment and Ranking of the Degree of Growth in Counseling

---

$W = .6674$ $\chi^2_r = 28.45$ 14 d.f. $P < .014228$

$\bar{r}' = .5155$

Clinical Rankings of Growth and Status Ratings of Adjustment

(N = 15) $r' = .582$ (N = 13) $r' = .643$

Lay Rankings of Growth and Status Ratings of Adjustment

(N = 15) $r' = .372$ (N = 13) $r' = .396$

Clinical Rankings of Growth and Lay Rankings of Growth

(N = 15) $r' = .627$ (N = 13) $r' = .687$

For an N of 14 and a one-tailed test:

$P = .05$ $.01$

$r' = .456$ $.645$

For an N of 16 and a one-tailed test:

$P = .05$ $.01$

$r' = .425$ $.601$

However, the relationship between the averaged criterion rankings and the D values between Levels I-A and I-B and between the averaged criterion rankings and the Z' values between Levels I-A and I-B should not be significant because these measures are, theoretically, from the same operationally defined personality level. Levels I-A and I-B are both other's ratings of the subjects, while Level III is ratings of projective test material. (See footnote 4 in this chapter.)

An inspection of Tables 16, 17, and 18, clearly indicates that, although the correlations between the criterion values and the post D and Z' values tend toward the correct direction for Levels I-A vs III and I-B vs III, they are not statistically significant. However, Table 18 shows that there is a significant positive relationship at or beyond the .05 level for four of the six correlations between the P.S.T. criterion rankings and the post D and Z' values between the Clinical and Lay raters. (Levels I-B and I-A.) This occurrence runs contrary to the theoretical expectation that there should not be a relationship between the criterion and the Level I-A and I-B discrepancies.

Although it is possible that this is, in fact, only a chance occurrence (Type I error, if the null hypothesis were rejected) a reasonable explanation may also be considered. The P.S.T. criterion rankings are based on Level III information, and the Clinical and Lay rating are based on Level I information. Thus, when a subject's projective stories were judged as being healthier, it is possible that his interpersonal behavior was consistent enough to be judged reliably and to produce, thus, a small D value and a large Z' value between Levels I-A and I-B. Conversely, when a subject's projective material was judged as being more emotionally disturbed, it is possible that his interpersonal behavior was markedly inconsistent and consequently resulted in a large D value

Table 16

Correlations Between the Averaged Criterion Rankings and the Discrepancy Between Levels I-A, I-B, and III on THE BEHAVIOR INVENTORY
(N = 13)

Criterion Rankings	Personality Levels	Pre or Post Rx.	r with D values: All Raters and Items	r with D values: All Raters but Some Items Dropped	r with D values: Some Raters and Items Dropped	r with Z' values: All Raters and Items	r with Z' values: All Raters but Some Items Dropped	r with Z' values: Some Raters and Items Dropped
Clinical Rankings	I-A vs. III i.e. Lay vs. P.S.T.	Pre	-.136	-.151	-.298	.284	.253	.145
		Post	.230	.318	.150	-.134	-.088	-.065
Lay Rankings	I-B vs. III Clinical vs. P.S.T.	Pre	-.144	-.282	-.164	.160	.101	.324
		Post	.362	.431	.104	-.253	-.215	-.166
P.S.T. Rankings	I-B vs. I-A Clinical vs. Lay Raters	Pre	.190	-.056	-.294	-.184	.231	.322
		Post	.234	.247	.149	-.447	-.329	-.298

I-A = Lay Raters on Kinescope
I-B = Clinical Raters on Kinescope
III = P.S.T. Raters

For 11 d.f. and a one-tailed test:

P =	.05	.025
r =	.476	.553

Table 17

Correlations Between the Averaged Criterion Rankings and the Discrepancy Between Levels I-A, I-B, and III on THE NEED BEHAVIOR SCALE
(N = 13)

			r with D values			r with Z' values		
Criterion Rankings	Personality Levels	Pre or Post Rx.	All Raters and Items	All Raters but Items Dropped	Both Raters and Items Dropped	All Raters and Items	All Raters but Items Dropped	Both Raters and Items Dropped
Clinical Rankings	I-A vs. III i.e. Lay vs. P.S.T.	Pre	.117	-.011	.033	.455	.452	.134
		Post	.233	.208	.391	-.112	-.266	-.457
Lay Rankings	I-B vs. III Clinical vs. P.S.T.	Pre	.141	.259	.276	.330	.120	.005
		Post	.166	.146	.527	-.047	-.237	-.510
P.S.T. Rankings	I-B vs. I-A Clinical vs. Lay Raters	Pre	-.067	-.136	-.320	.227	.376	.476
		Post	.190	.178	.299	-.172	-.040	.051

I-A = Lay Raters on Kinescope
I-B = Clinical Raters on Kinescope
III = P.S.T. Raters

For 11 d.f. and a one-tailed test:

P =	.05	.025
r =	.476	.553

Table 18

Correlations Between the Averaged Criterion Rankings and the Discrepancy Between Levels I-A, I-B, and III on THE BEHAVIOR RATING SCALE (N = 13)

			r with D values			r with Z' values		
Criterion Rankings	Personality Levels	Pre or Post Rx.	All Raters and Items	All Raters but Items Dropped	Both Raters and Items Dropped	All Raters and Items	All Raters but Items Dropped	Both Raters and Items Dropped
Clinical Rankings	I-A vs. III i.e. Lay vs. P.S.T.	Pre	.036	.155	.074	.354	.218	.214
		Post	.340	.373	.370	-.404	-.302	-.458
Lay Rankings	I-B vs. III Clinical vs. P.S.T.	Pre	.234	.255	.205	-.075	.094	.104
		Post	.299	.165	.550	-.277	-.164	-.468
P.S.T. Rankings	I-B vs. I-A Clinical vs. Lay Raters	Pre	.166	.099	.253	-.271	-.175	-.254
		Post	.486	.433	.464	-.574	-.486	-.399

I-A = Lay Raters on Kinescope
I-B = Clinical Raters on Kinescope
III = P.S.T. Raters

For 11 d.f. and a one-tailed test:

P =	.05	.025
r =	.476	.553

and a small Z' value.[10] This interaction could result in the significant correlations between the P.S.T. rankings and the D and Z' values between Levels I-B and I-A.

The Behavior Inventory was used as a part of the original study, and from it pre and post counseling self ratings by the subjects were obtained. This material then represents Level II or Leary's Conscious Descriptions (Leary, 1957, 78-79). Therefore, it was possible to test the first hypothesis in relation to discrepancies between Level II and Levels I-B, I-A, and III, and the criterion. These results are presented in Table 19. Only one of the thirty-six correlations with the post values was significant. It is likely that this would also be a Type I error if the null hypothesis were rejected on the basis of only one significant correlation.

Although a number of the correlations between the criterion and the post D and Z' values were in the predicted direction, only one was significant and it was not in the predicted direction. Thus the null hypothesis cannot be rejected. This does not "disprove" the first hypothesis, but rather leaves the problem unanswered. In terms of this study one cannot state what the nature of the relationship is between success in group counseling and the discrepancies between levels of personality at the conclusion of counseling.

Testing Hypothesis II

The second hypothesis predicted that the greater the degree of success in group counseling, the greater the reduction in the discrepancies between levels of personality from pre to post counseling. This hypothesis was tested

---

[10] The small D and a large Z' value were considered to represent a small discrepancy between Levels of personality. The criterion rankings were also scored so that a small score indicated improvement and a large value indicated lack of improvement. Thus the desirable correlation between criterion and the D values should be positive and between the criterion and Z' values should be negative.

Table 19

Correlations Between the Criterion Rankings and the Discrepancy Between Levels I-A, I-B, III and the Self Ratings on THE BEHAVIOR INVENTORY

Criterion Rankings	Personality Levels		Pre or Post	r with D values: All Raters and Items	r with D values: All Raters but Some Items Dropped	r with D values: Some Raters and Items Dropped	r with Z' values: All Raters and Items	r with Z' values: All Raters but Some Items Dropped	r with Z' values: Some Raters and Items Dropped	N
Clinical Rankings	Lay vs. Self	I-A vs. II	Pre	-.334	-.329	-.325	.168	.194	.206	14
			Post	.206	.216	.192	.101	.108	.364	
Clinical Rankings	P.S.T. vs. Self	III vs. II	Pre	-.344	-.340	-.331	-.090	-.143	-.149	13
			Post	.243	.264	.251	-.054	-.203	-.063	
Lay Rankings	Clin. vs. Self	I-B vs. II	Pre	-.150	-.138	-.276	.390	.351	.370	14
			Post	.381	.324	.322	.551	.414	.380	
Lay Rankings	P.S.T. vs. Self	III vs. II	Pre	-.172	-.172	-.169	-.192	-.214	-.078	13
			Post	.424	.444	.421	.161	.306	.206	
P.S.T. Rankings	Lay vs. Self	I-A vs. II	Pre	-.016	-.013	-.014	-.265	-.255	-.272	13
			Post	-.101	-.091	-.094	-.310	-.244	-.221	
P.S.T. Rankings	Clin. vs. Self	I-B vs. II	Pre	-.007	-.009	.040	-.234	-.248	-.253	13
			Post	.203	-.143	-.127	.182	.266	.094	

I-A = Lay Raters on Kinescope
I-B = Clinical Raters on Kinescope
II = Self Ratings
III = P.S.T. Ratings

by obtaining a difference score between the pre and post D values and then correlating this with the criterion values. The post D values were algebraically subtracted from the pre D values because a plus difference would, theoretically, indicate improvement and a negative difference would be an indication of the subject's being more maladjusted. That is, it was predicted that the larger positive "pre minus post differences" would be associated with the criterion rankings of improvement and that small positive or negative "pre minus post differences" would be associated with criterion rankings indicating little or no improvement.

All of the correlations between criterion and difference scores were obtained by using the Spearman ranks method because the correlations were all "hand computed." Since both the Pearson product-moment formula and the Spearman are applicable to small samples, choice is only a matter of convenience. Both of the formulas ". . . are merely different algebraic forms of the same thing, so that, wherever the one is applicable, the other is also applicable" (Peters and Van Voorhis, 1940, 108-109).

Two different criterion measures were related to the difference scores. First, the status ratings of adjustment were correlated with the difference scores (see Tables 20, 21, and 22), and then the growth criterion rankings previously used were also correlated with the difference scores where confounding would not occur (see Table 23).

The use of the status ratings of adjustment as a criterion is not presented as a test of the second hypothesis, but rather as additional information concerning the relationship between an over-all rating of adjustment and the change in the discrepancies from pre to post counseling.

Of the thirty-six correlations in Tables 20, 21, and 22, only one is significant at or beyond the .05 level. It again appears that this one significant correlation could have occurred as a chance happening, and it would

Table 20

Rank Order Correlations Between the Status Ratings of Adjustment and the Pre Minus Post Differences of the D Values for THE BEHAVIOR INVENTORY

Personality Levels	All Raters and Items	All Raters but Some Items Dropped	Some Raters and Items Dropped	N
Clinical (I-B) vs. Lay Raters (I-A)	+.086	+.235	+.054	15
Clinical (I-B) vs. P.S.T. Raters (III)	+.080	+.132	-.011	13
Lay Raters (I-A) vs. P.S.T. Raters (III)	+.006	+.126	+.209	13
Clinical (I-B) vs. Self Ratings (II)	-.037	-.138	-.380	15
Lay Raters (I-A) vs. Self Ratings (II)	-.094	-.094	-.125	13
P.S.T. Raters (III) vs. Self Ratings (II)	+.110	+.143	+.143	13

Critical Values of r'

N	Significance Level (one-tailed test) .05	.01
12	.506	.712
14	.456	.645
16	.425	.601

Table 21

Rank Order Correlations Between the Status Ratings of Adjustment and the Pre Minus Post Differences of the D Values for THE NEED BEHAVIOR SCALE

Personality Levels	All Raters and Items	All Raters but Some Items Dropped	Some Raters and Items Dropped	N
Clinical (I-B) vs. Lay Raters (I-A)	-.078	-.175	+.339	15
Clinical (I-B) vs. P.S.T. Raters (III)	-.308	-.082	+.039	13
Lay Raters (I-A) vs. P.S.T. Raters (III)	-.093	+.033	+.308	13

Table 22

Rank Order Correlations Between the Status Ratings of Adjustment and the Pre Minus Post Differences of the D Values for THE BEHAVIOR RATING SCALE

Personality Levels	All Raters and Items	All Raters but Some Items Dropped	Some Raters and Items Dropped	N
Clinical (I-B) vs. Lay Raters (I-A)	+.129	+.154	+.493	15
Clinical (I-B) vs. P.S.T. Raters (III)	-.060	+.039	+.291	13
Lay Raters (I-A) vs. P.S.T. Raters (III)	-.275	-.176	+.110	13

be a Type I Error if the null hypothesis were rejected. Thus it seems clear that, in terms of this study, a relationship could not be demonstrated between a status measure of adjustment and the reduction of the discrepancies between the operationally defined levels of personality from pre to post counseling. The two variables, the status ratings of adjustment and the "pre minus post differences," appear to be unrelated.

The second hypothesis was tested by correlating the growth criterion scores with the "pre minus post differences" using D values generated by combinations of Lay, Clinical, and P.S.T. raters which would not be confounded with the criterion rankings by the Lay and Clinical groups. Thus only the correlations between the Clinical rankings of growth and the difference scores obtained from the original Lay vs. P.S.T. item D values were presented for interpretation. Also, only the correlations between the Lay rankings of growth and the difference scores obtained from the original Clinical vs. P.S.T. item D values were considered for interpretation.

An inspection of Table 23 shows that the correlations between the criterion and the difference scores from the Need Behavior Scale and the Behavior Rating Scale are generally not significant. Of the twelve possible correlations, only one even approaches significance.[11] However, three of the six correlations between the criterion and the "pre minus post difference" scores from The Behavior Inventory are significant at or beyond the .05 level and are in the predicted directon. The remaining three correlations are also in the predicted direction with the largest one slightly below the .05 level of significance. Thus it would appear statistically reasonable to reject the null hypothesis and to accept the hypothesis that the greater the degree of

---

[11] See Table 20 for the critical values of the Spearman Rank Correlation Coefficient.

Table 23

Rank Order Correlations Between the Criterion Rankings and the Pre Minus Post Differences (N = 13)

		r' With Difference Score		
Criterion	Source of the Original D Values	All Raters and Items Used	All Raters but Some Items Dropped	Some Raters and Items Dropped
THE BEHAVIOR INVENTORY				
Clinical Rankings of Growth	Lay Raters and P.S.T. Raters	.291	.456	.571
Lay Rankings of Growth	Clinical Raters and P.S.T. Raters	.506	.506	.077
NEED BEHAVIOR SCALE				
Clinical Rankings of Growth	Lay Raters and P.S.T. Raters	.064	.159	.484
Lay Rankings of Growth	Clinical Raters and P.S.T. Raters	-.176	-.126	.256
BEHAVIOR RATING SCALE				
Clinical Rankings of Growth	Lay Raters and P.S.T. Raters	.044	.148	.242
Lay Rankings of Growth	Clinical Raters and P.S.T. Raters	-.055	-.214	.264

success in group counseling, the greater the reduction in the discrepancies between levels of personality from pre to post counseling. However, further analysis raises some serious methodological problems that indicate that this correlation is not an accurate measure of change.

The first limitation is that this hypothesis could only be accepted in terms of the relationship between the criterion of growth and the reduction in the discrepancies between Levels I-A (Lay raters) and III (P.S.T. raters) and between Levels I-B (Clinical raters) and III (P.S.T. raters). It could not be applied to differences between Level II (self ratings) and III (P.S.T. ratings), nor to differences between Level II (self ratings) and I-A (Lay raters) or I-B (Clinical raters). A second limitation must be applied in connection with the type of items used in The Behavior Inventory. A third limitation, also to be discussed below, must be considered because further study of the data raised some methodological questions concerning the nature of the correlation between the criterion and the apparent pre to post reduction in the discrepancies between Level I and III.

The items on The Behavior Inventory are weighted in the direction of describing behavior in general terms of action and the subjects approach to others. It is possible that a contrast between this type of description of behavior in the interpersonal situation of group counseling and the stories of a T.A.T. type of projective instrument (difference between Levels I and III) might have some relationship to the more general clinical type of judgment of improvement in counseling. It is simply suggested that the nature of these items in terms of their describing activity in an interpersonal situation (be it a counseling group or a T.A.T. type of story), might prove to be a significant factor in further research.

The means of the post item D values should be consistently smaller than the means of the pre item D values because the prediction is that the greater

Table 24

Pre and Post Means For Item D Values and Pre Minus Post Mean Differences
(N = 13)

Source of the Original Item D Values	Pre or Post	All Raters and Items		All Raters but Some Items Dropped		Some Raters and Items Dropped	
		Mean	Pre-Post Diff.	Mean	Pre-Post Diff.	Mean	Pre-Post Diff.
THE BEHAVIOR INVENTORY							
Lay Raters and P.S.T. Raters	Pre	.755	(-.135)	.565	(-.186)	.639	(-.131)
	Post	.890	t=1.2617	.751	t=2.0284	.770	t=1.2018
Clinical Raters and P.S.T. Raters	Pre	.865	(-.073)	.675	(-.073)	.765	(-.053)
	Post	.938	t=1.1515	.748	t=0.6239	.818	t=0.5048
NEED BEHAVIOR SCALE							
Lay Raters and P.S.T. Raters	Pre	5.47	(0.20)	4.20	(0.06)	5.19	(-0.11)
	Post	5.27	t=0.4464	4.14	t=0.1714	5.30	t=0.2895
Clinical Raters and P.S.T. Raters	Pre	5.78	(-0.31)	4.56	(-0.04)	5.24	(0.16)
	Post	6.09	t=0.5400	4.60	t=0.0756	5.08	t=0.2822
BEHAVIOR RATING SCALE							
Lay Raters and P.S.T. Raters	Pre	7.73	(0.26)	5.48	(0.28)	6.84	(-0.10)
	Post	7.47	t=0.3698	5.20	t=0.3655	6.94	t=0.1603
Clinical Raters and P.S.T. Raters	Pre	8.58	(0.09)	6.39	(-0.13)	7.31	(-0.60)
	Post	8.49	t=0.1355	6.52	t=0.2163	7.91	t=0.8510

For a one-tailed test and 12 d.f.

t = 2.179	p = .025	t = 1.083	p = .15
t = 1.782	p = .05	t = 0.876	p = .20
t = 1.356	p = .10		

the success, the greater the reduction in the discrepancy measures (i.e., the item D values) from pre to post counseling. The means of the pre and post item D values and the "pre minus post mean differences" shown in Table 24 have not followed this expectation; rather the post means have tended to be larger. All of the means for the post item D values were larger than the means for the pre item D values on the data taken from The Behavior Inventory ratings.

A series of "t" tests for the difference between the pre and post means of the item D values were computed using the formula for the standard error of the difference between the means based on paired or matched subjects (Edwards, 1954, 278-282). The use of this formula would actually be a slight statistical disadvantage for the testing of the difference between the pre and post means for the item D values from the Lay and P.S.T. raters on The Behavior Inventory as shown by the negative correlations in Table 25. However, it was judged to be the most appropriate standard error formula for the majority of the "t" tests in Table 24. The "t" test procedure is an appropriate method to test the difference between the means even when the data is in the form of difference scores (Walker and Lev, 1953, 158-159). Therefore, it was decided that the same procedure could be applied to the difference between the means of the item D values.

The tendency for the post item D value means to be larger than the pre means for the data from The Behavior Inventory (see Table 24) is accorded some statistical support by the fact that one of the six "t" tests is significant beyond the .05 level and that three more are between the .15 and .10 levels of significance. Although this result is not conclusive evidence that the post item D values are significantly larger, it can be interpreted as evidence that there was a rather marked tendency in that direction. In addition, it should be noted that only one of the twelve "t" tests on the data from the Need

Table 25

Correlation Between the Item D Values Pre and the Item D Values Post (N = 13)

Item D Between	All Raters and Items	All Raters but Some Items Dropped	Some Raters and Items Dropped
THE BEHAVIOR INVENTORY			
Lay Raters and P.S.T. Raters	-.150	-.064	-.295
Clinical Raters and P.S.T. Raters	.251	.015	.310
NEED BEHAVIOR SCALE			
Lay Raters and P.S.T. Raters	.216	.516	.627
Clinical Raters and P.S.T. Raters	.054	.304	.385
BEHAVIOR RATING SCALE			
Lay Raters and P.S.T. Raters	.210	-.078	.353
Clinical Raters and P.S.T. Raters	.297	.299	.241

Behavior Scale and Behavior Rating Scale even approached the .20 level of significance.

When the "pre item D minus post item D difference" scores for the data from The Behavior Inventory are inspected (Tables 26 and 27), it is again apparent that the nature of the change from pre to post counseling was not in the expected direction. That is, the majority of the "pre minus post difference" scores are negative. This also indicates that the post item D values were, on the average, larger than the pre item D values. With this additional information, the interpretation of the significant correlations in Table 23, becomes questionable, and it is apparent that one cannot accept the hypothesis that the greater the degree of success in group counseling, the greater the reduction in the discrepancies between Levels (I and III) of personality from pre to post counseling.

This relationship can be summarized in the following manner: there was a marked tendency for the post item D values to be larger than the pre item D values. In terms of the original theoretical formulations this tendency would indicate that the subjects became more maladjusted. Although there is this suggestion of increased maladjustment, yet the growth criterion demonstrates a significant positive relationship with the "pre minus post difference" scores from The Behavior Inventory data. This is an unexpected result but perhaps it can be explained by the fact that only two of the original four groups (the ones for which kinescopes had been made) were used in this study, and this included the only group out of the four for which previous investigators had reported no significant growth (Broedel, Ohlsen, Proff, and Southard, 1960). However, there are several other possible explanations or interpretations as to what could have brought this about.

First, there is the possibility that the criterion is in error and, in fact, all of the subjects did become worse in terms of their mental health adjustment. This conclusion is to be doubted because the clinicians making the

Table 26

Rank Order of Growth and the Difference Between Pre and Post D Values For the Lay vs. P.S.T. Ratings on THE BEHAVIOR INVENTORY

Subject Number	Rank Order of Growth (1 = Most - 13 = Least) (Clinical Criterion Ranks)	Amount of Pre Minus Post Difference		
		All Raters and Items	All Raters but Some Items Dropped	Some Raters and Items Dropped
11	1	-.220	-.236	-.074
5	2	-.454	-.276	-.173
1	3	-.214	-.180	-.067
13	4	-.325	-.242	-.146
9	5	.089	-.088	-.173
2	6	.474	.102	.033
6	7	-.029	.058	.090
10	8	.718	.662	.924
12	9	.014	-.201	-.203
7	10	-.373	-.449	-.571
15	11	-.293	-.370	-.245
8	12	-.522	-.545	-.717
14	13	-.628	-.652	-.456

Table 27

Rank Order of Growth and the Difference Between Pre and Post D Values
For the Clinical vs. P.S.T. Ratings on
THE BEHAVIOR INVENTORY

Subject Number	Rank Order of Growth (1 = Most - 13 = Least) (Lay Criterion Ranks)	Amount of Pre Minus Post Difference		
		All Raters and Items	All Raters but Some Items Dropped	Some Raters and Items Dropped
2	1	.039	.010	-.124
11	2	.037	.043	.133
13	3	-.061	.187	-.075
5	4	-.455	-.175	-.400
12	5	.795	.793	.658
9	6	.436	.223	.253
7	7	-.310	-.423	-.275
6	8	-.119	-.029	.056
10	9	.305	.238	.239
1	10	-.453	-.415	-.391
8	11	-.375	-.863	-.785
14	12	-.568	-.479	-.264
15	13	-.326	-.062	.281

Clinical rankings of growth were well trained and demonstrated significant agreement among themselves (see Table 9) and their rankings also demonstrated a significant relationship with the Lay rankings of growth (see Tables 13 and 14). However, all of clinicians had had limited experience with "normal" adolescents. The significant correlation between the Clinical and Lay rankings of mental health growth also indicates that changes in the subjects' behavior during the counseling did occur and that these changes were reasonably apparent and observable by both trained Clinical observers and by Lay observers. Thus, the criterion is considered valid, and, further, it is concluded that the subjects were successfully ranked according to their improvement in mental health. At the conclusion of the data collection from the Clinical and Lay raters, several of the raters made informal comments to the investigator indicating that they thought that the subjects had shown considerable improvement.

A second possible source of difficulty is that the two rating scales were not sensitive enough measures of the subjects' behavior. The generally negative results obtained indicate that the use of Osgood's Semantic Differential (as used in Aaron Smith's thesis) as well as The Behavior Inventory, might have provided a more suitable choice of instruments (Osgood, 1957; Smith, 1958, 25-26). It is now suggested that the dimensions of meaning as measured by the Semantic Differential may be a proper choice for further research concerning "Levels of Personality."

A third possible source of error is that the subjects were not sufficiently neurotic, but rather were personality or character disorders. That is, that they were not sufficiently anxious and were not sufficiently distressed by their difficulties. To the degree that the subjects were experiencing only minimal subjective anxiety with only a slight sense of distress, the formulations developed in the first chapter would not apply. However, a previous study by Katz, et al., using a conceptual approach which was similar to Horney's theory of neurosis, resulted in significant findings (Ohlsen and Proff, 1960 69-81).

Thus, one might expect the subjects to have been sufficiently neurotic-like to be used in this study.

A fourth possibility is that, though the subjects did demonstrate improvement in the group setting, they did not demonstrate improvement in terms of the Picture Story Test material as rated in this study. Such a phenomenon could produce the increase in the post item D values. In a more general sense, it is possible that the types of behavior demonstrated in the counseling groups and in the Picture Story Tests became more discrepant at the conclusion of counseling. Again some of the comments of the P.S.T. raters indicated that their general impression was that the projective material they rated appeared "sicker" or more maladjusted in its content. This could explain the increase in the post item D values and also the near zero and negative correlations between the P.S.T. criterion rankings of growth with the Clinical and Lay rankings of growth (see Tables 13, 14, and 24). However, this cannot explain the positive and significant correlations between the Clinical and Lay criterion rankings and "the pre minus post difference" scores (see Table 23).

The final possible source of error is that the original formulation was in error and that the relationship between success in group counseling and the reduction in the discrepancies between levels of personality from pre to post counseling may not be linear, but rather curvilinear. An inspection of Table 26 clearly suggests this possibility because both the subjects judged to have grown most and those judged to have grown least have negative "pre minus post difference" scores, while the subjects between the extremes on the growth criterion generally have positive difference scores. This indicates that both the subjects who grew most and those who grew least had increases in the discrepancies between personality Levels I and III, while those judged to have been intermediate in mental health growth actually did decrease in their discrepancies. However, this relationship is not apparent in the data presented in Table 27. It is suggested that future research concerning the relationships between success in counseling or psychotherapy and the discrepancies between levels of personality include the hypotheses of curvilinearity. In any case the small N requires that these relationships must be regarded as highly tentative.

Summary

The major premise of the study has been that there is an inverse relationship between success in group counseling and the discrepancy between levels of personality. The discrepancies between levels of personality were measured by the item D procedure and the product-moment correlations which were converted to Z' values. These discrepancy scores or values were then related to the criterion of success.

The first hypothesis predicted that the greater the degree of success in group counseling, the smaller the discrepancy between levels of personality at the conclusion of counseling. The criterion of success in group counseling was not significantly related to the discrepancy between levels of personality at the conclusion of counseling.

The second hypothesis predicted that the greater the degree of success in group counseling, the greater the reduction in the discrepancies between levels of personality from pre to post counseling. The criterion of success in group counseling was significantly related to the "pre minus post difference" scores for the discrepancy between personality Levels I and III (*i.e.*, Kinescope ratings and the ratings of the Picture Story Test stories). However, a series of "t" tests for the difference between the pre and post means of the item D values and an inspection of the difference scores indicated that there had not been a reduction in the discrepancies between Levels I and III from pre to post counseling.

Thus neither of the null-hypotheses could be properly rejected.

Several explanations were offered concerning the negative results of this study and it was concluded that future research concerning the relationships between success in counseling or psychotherapy and the discrepancies between levels of personality include a hypothesis of curvilinearity.

CHAPTER VI

SUMMARY AND CONCLUSIONS

Summary

There are many theoretical statements which postulate a relationship between emotional maladjustment and interlevel personality conflicts. The implications common in counseling and clinical lore are that the reduction in these conflicts will result in improvement in mental health. Many research findings indicate that the induction of conflictual situations in both the lower animals and human subjects results in maladjustive behavior. Other studies suggest that for human subjects, when conflict is resolved and integration is achieved, there is a reduction in the discrepant levels of meaning or behavior. Previous studies also indicate that levels of personality can be operationally defined and measured.

However, an unpublished doctoral dissertation, which predicted that the amount of interlevel conflict or discrepancy in the meaning of self is positively and significantly related to the degree of emotional disturbance in persons, concluded that the assumption that various levels of personality were represented in the tests was rejected because there was an alpha coefficient of .85 for six of the twelve variables. This result was interpreted as meaning that these six variables did not represent independent dimensions and were not drawn from different levels of personality (Smith, 1958). A factor analysis of the data from Smith's dissertation indicated that the assumption based on the alpha coefficient was in error. This was interpreted as further evidence that there are, in fact, different levels of personality and that they can be defined and measured.

It was thus hypothesized that:

I. The greater the degree of success in group counseling, the smaller the discrepancy between levels of personality at the conclusion of counseling; and

II. The greater the degree of success in group counseling, the greater the reduction in the discrepancies between levels of personality from pre to post counseling.

The following procedure was used to test these two hypotheses. Two sets of kinescope recordings selected from the beginning and ending sessions of two counseling groups were viewed and the subjects independently rated by five clinical psychologists and five secretaries. The pre and post sets of stories written to five T.A.T. type of pictures were studied and independently rated by four clinical psychologists. These ratings were made on three different rating scales.

Operational definitions of levels of personality based on the sources of the data were used (Leary, 1957). The levels were defined as follows: Level I is other's ratings of the subject's "public behavior"; Level II, self ratings; and Level III, the ratings of the subject's fantasy productions.

The technique of combining ratings of several judges in order to increase reliability, a procedure the validity of which has been confirmed by a recent experiment, has been used throughout this study. Rater reliability computed across items on subjects and across subjects on items for both the pre and post ratings demonstrated a moderate, but sufficient agreement among the raters. A low to moderate interrelationship was found for the item D values between selected individual raters. This relationship was judged to be sufficient for the item D reliabilities, which usually are low in terms of their absolute values because they are a reliability of a difference source.

Criterion measures of success in group counseling were obtained from several different sources and were tested for reliability and then related to measures of differences between the subject's levels of personality. This last operation formed the basis of the hypothesis testing.

Conclusions

Analysis of Smith's data failed to find one general factor. It did find four factors which could be interpreted in terms of Leary's definitions of levels of personality.

Factors I and IV were interpreted as being a measure of personality Level II, or Conscious Self Description.

Factor III was interpreted as being a measure of personality Level III, or Private Symbolization.

Factor II, a bipolar factor, was interpreted as representing aspects of both personality Level II and Level III and was named Neurotic Defensiveness.

This analysis suggests that future attempts to measure or use "levels of personality" should follow Leary's operational definitions.

The analysis of the data from the kinescope recordings, T.A.T. type stories, and the criterion of success in group counseling found that:

The criterion of success in group counseling was not significantly related to the discrepancy between levels of personality at the conclusion of counseling.

The criterion of success in group counseling was significantly related to the "pre minus post difference" scores for the discrepancy between personality Levels I and III (*i.e.*, Kinescope ratings and the ratings of the Picture Story Test stories). However, a series of "t" tests for the difference between the pre and post means of the item D values and an inspection of the difference scores indicated that there had not been a reduction in the discrepancies between Levels I and III from pre to post counseling.

Thus neither of the null-hypotheses could properly be rejected.

Several explanations were offered concerning the negative results of this study, and it was concluded that future research concerning the relationships between success in counseling or psychotherapy and the discrepancies between levels of personality include a hypothesis of curvilinearity. In any case the small sample (N) requires that these relationships be regarded as highly tentative.

REFERENCES

Adler, A. 1917. *The neurotic constitution*. New York: Moffat, Yard and Company.

Adler, A. 1929. *The practice and theory of individual psychology*. London: Routledge and Kegan Paul, Ltd.

Arieti, S. 1955. *Interpretation of schizophrenia*. New York: Robert Brunner.

Bales, R. F. 1950. *Interaction process analysis*. Cambridge: Addison-Wesley Press.

Bellak, L. ed. 1958. *Schizophrenia: a review of the syndrome*. New York: Logos Press.

Brenman, M. and Gill, M. M. 1947. *Hypnotherapy*. New York: International Universities Press.

Brickner, R. M. and Kubie, L. S. 1936. A minature psychotic storm produced by superego conflict over simple posthypnotic suggestion. *Psychoanalytic Quart*. 11. 467.

Broedel, J. W. 1958. A study of the effects of group counseling on the academic performance and mental health of underachieving gifted adolescents. Unpublished Doctoral Thesis. Urbana: University of Illinois.

Broedel, J., Ohlsen, M., Proff, F. and Southard, C. The effects of group counseling on gifted underachieving adolescents. *Jn. Coun Psychol*. 7. 163-170.

Cronbach, Lee J. 1949. *Essentials of psychological testing*. New York: Harper and Brothers.

Cronbach, Lee J. 1960. *Essentials of psychological testing*. New York: Harper and Brothers.

Diamond, Solomon. 1959. *Information and error*. New York: Basic Books.

Dickman, Kern W. 1960. Factorial validity of a rating instrument. Unpublished Doctoral Thesis. Urbana: University of Illinois.

Dollard, J. and Miller, N. E. 1950. *Personality and psychotherapy*. New York: McGraw-Hill.

Edwards, Allen L. 1950. *Experimental design in psychological research*. New York: Rinehart and Company.

Edwards, Allen L. 1954. *Statistical methods for the behavioral sciences*. New York: Rinehart and Company.

Erickson, M. H. 1935. A study of an experimental neurosis hypnotically indiced in a case of ejaculatio precox. *Brit*. *Jn*. *Med*. *Psychol*. 15. (Part I) 34-50.

Erickson, M. H. 1939. An experimental demonstration of the psychopathology of everyday life. *Psychoanal*. *Quart*. 8. 338-353.

Erickson, E. H. 1950. *Childhood and society*. New York: W. W. Norton.

Fiedler, Fred E., *et al*. 1958. *Quasi-therapeutic relations in small college and military groups*. Group Effectiveness Research Laboratory. Department of Psychology. Urbana: University of Illinois.

Freedman, M., *et al*. 1951. *The interpersonal dimension of personality*. *Jn*. *Pers*. 20. 143-161.

Freeman, G. L. 1940. A method of inducing frustration in human subjects and its influence upon palmar skin resistance. *Amer*. *Jn*. *Psychol*. 53. 117-120.

Freud, S. 1931. *The interpretation of dreams*. New York: Carlton House.

Freud, S. 1957. *Collected papers*. Vol. II. London: Hogarth Press.

Gardner, Lindzey. 1954. *Handbook of social psychology*: *Theory and method*. Vol. I. Cambridge: Addison-Wesley.

Gazda, George M. 1959. The effects of short term group counseling on prospective counselors. Unpublished Doctoral Thesis. Urbana: University of Illinois.

Grummon, D. L. and Butler, J. M. 1953. Another failure to replicate Keet's study "Two verbal techniques in a miniature counseling situation." *Jn*. *Abnorm*. *Soc*. *Psychol*. 48. 4. 597.

Hall, C. S. and Lindzey, G. 1957. *Theories of personality*. New York: John Wiley and Sons.

Harman, Harry H. 1960. *Modern factor analysis*. Chicago: The University of Chicago Press.

Hebb, D. O. 1949. *The organization of behavior*. New York: John Wiley and Sons.

Horney, K. 1937. *The neurotic personality of our time*. New York: W. W. Norton.

Horney, K. 1945. *Our inner conflicts*. New York: W. W. Norton.

Hovland, C. I. and Sears, R. R. 1938. Experiments on motor conflict. I. Types of conflict and their modes of resolution. *Jn*. *Exp*. *Psychol*. 23. 477-493.

Hunt, J. McV. ed. 1944. *Personality and the behavior disorders*. Vol. I. New York: Ronald Press.

Johoda, Marie. 1958. *Current concepts of positive mental health.* New York: Basic Books.

Jahoda, M., *et al*. *Research methods in social relations*. *Basic processes.* Part I. New York: The Dryden Press.

Johnson, D. M. and Vidulich, R. N. 1956. Experimental manipulation of the halo effect. *Journal of Applied Psychology*. 40. 2. 130-134.

Jones, E. 1953. *The life and works of Sigmund Freud*. Vol. I. New York: Basic Books.

Jung, C. G. 1953. *Two essays on analytical psychology*. New York: Pantheon Books.

Kaiser, Henry F. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*. 23. 3. 187-200.

Kaiser, Henry F. 1960. Comments on communalities and the number of factors (a dittoed paper). Bureau of Educational Research. Urbana: University of Illinois.

Keet, C. D. 1948. Two verbal techniques in a miniature counseling situation. *Psychological Monographs*: *General and Applied*. 62. 7. 1-55.

Krasnogorski, N. I. 1925. The conditioned reflexes and children's neuroses. *Amer*. *Jn*. *Dis*. *Child*. 30. 753-786.

LaForge, R., *et al*. 1954. The interpersonal dimension of personality: II An objective study of repression. *Jn*. *Pers*. 23. 1. 129-153.

LaForge, R. and Suczek, R. F. 1955. The interpersonal dimension of personality: III. An interpersonal check list. *Jn*. *Pers*. 24. 1. 94-112.

Leary, T. 1956. A method and methodology for measuring fantasy and imaginative expression. *Jn*. *Pers*. 25. 2. 159-175.

Leary, T. 1957. *Interpersonal diagnosis of personality*. New York: Ronald Press.

Leary, T. and Coffey, H. 1954. The prediction of interpersonal behavior in group psychotherapy. *Group Psychother*. 7. 7-51.

Leary, T. and Coffey, H. 1955. Interpersonal Diagnosis: Some problems of methodology and validation. *Jn*. *Abnorm*. *Soc*. *Psychol*. 50. 110-124.

Leary, T. and Harvey, J. 1956. A methodology for measuring personality changes in psychotherapy. *Jn*. *Clin*. *Psychol*. 12. 123-132.

LeCron, L. M. 1954. *Experimental hypnosis*. New York: Macmillan Company.

Lindquist, E. F., *et al*. 1951. *Educational measurement*. Washington, D. C.: American Council on Education.

Lorr, Maurice, *et al*. 1960. Inpatient multidimensional psychiatric scale. a *restricted* experimental rating scale, form (July, 1960) no. 3 . Washington, D. C.: Central Outpatient Psychiatric Research Laboratory.

McDowell, M. 1948. An abrupt cessation of major neurotic symptoms following an hypnotically induced artificial conflict. *Bull. Menn. Cl.* 12. 168-177.

McNemar, Quinn. 1955. *Psychological statistics*. New York: John Wiley and Sons.

Maier, N. R. F. 1949. *Frustration: The study of behavior without a goal*. New York: McGraw-Hill.

Maslow, A. H. 1954. *Motivation and personality*. New York: Harper and Brothers.

Masserman, J. H. 1943. *Behavior and neurosis*. Chicago: University of Chicago Press.

Masserman, J. H. 1955. *The practice of dynamic psychiatry*. Philadelphia: W. B. Saunders.

Moss, C. S. 1953. A quantitative semantic analysis of the dream in psychotherapy. Unpublished Doctoral Thesis. Urbana: University of Illinois.

Mowrer, O. H. 1950. *Learning theory and personality dynamics*. New York: Ronald Press.

Mullahy, P. 1955. *Oedipus myth and complex*. New York: Grove Press.

Munroe, R. L. 1955. *Schools of psychoanalytic thought*. New York: Dryden Press.

Noble, Frank C. 1958. A method for the quantification of interaction in psychotherapeutic groups. Unpublished Doctoral Thesis. Urbana: University of Illinois.

Ohlsen, M. M. and Proff, F. C. 1960. *The extent to which group counseling improves the academic and personal adjustment of underachieving gifted adolescents*. Urbana: College of Education, University of Illinois.

Ohlsen, M. M., Proff, F. C., and Southard, C. 1959. The effects of group counseling on two groups of gifted adolescent underachievers. (a dittoed paper). Urbana: College of Education, University of Illinois.

Osgood, Charles E., *et al*. 1957. *The measurement of meaning*. Urbana: University of Illinois Press.

Pavlov, I. P. 1927. *Conditioned reflexes*. Humphrey Milford: Oxford University Press.

Peters, C. C. and Van Voorhis, W. R. 1940. *Statistical procedures and their mathematical basis*. New York: McGraw-Hill.

Rank, O. 1929. *The trauma of birth*. New York: Harcourt, Brace and Company.

Rank, O. 1950. *Will therapy and truth and reality*. New York: Alfred A. Knopf.

Razram, G. H. S. 1933. Conditioned responses in children. No. 148. New York: *Arch. Psychol*.

Reyher, J. 1958. Hypnotically induced conflict in relation to subception, repression, antisocial behavior and psychosomatic reactions. Unpublished Doctoral Thesis. Urbana: University of Illinois.

Sears, R. R. and Hovland, C. I. 1941. Experiments on motor conflict. II. Determination of mode of resolution by comparative strengths of conflicting responses. *Jn. Exp. Psychol*. 28. 280-286.

Shaffer, L. F. and Shoben, E. J., Jr. 1956. *The psychology of adjustment*. Boston: Houghton Mifflin.

Siegel, Sidney. 1956. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Smith, A. 1958. Discrepancy in the meaning of self in a multilevel personality system and emotional disturbance. Unpublished Doctoral Thesis. Urbana: University of Illinois.

Stein, Morris I. 1955. *The thematic apperception test*. Cambridge: Addison-Wesley.

Stern, G. G., Stein, M. I., and Bloom, B. S. 1956. *Methods in personality assessment*. Glencoe, Illinois: Free Press.

Sullivan, H. S. 1953. *The interpersonal theory of psychiatry*. New York: W. W. Norton.

Sullivan, H. S. 1956. *Clinical studies in psychiatry*. New York: W. W. Norton.

Thompson, C. 1950. *Psychoanalysis: evolution and development*. New York: Hermitage House.

Thorndike, Edward L. 1920. A constant error in psychological ratings. *Jn. of Applied Psychology*. IV. 25-29.

Thurstone, L. L. 1947. *Multiple-factor analysis*. Chicago: The University of Chicago Press.

Walker, H. M. and Lev, Joseph. 1953. *Statistical inference*. New York: Henry Holt and Company.

Watkins, J. G. 1949. *Hypnotherapy of war neuroses*. New York: Ronald Press.

Way, L. 1950. *Adler's place in psychology*. New York: The Macmillan Company.

Weitzenhoffer, A. M. 1957. *General techniques of hypnotism*. New York: Grune and Stratton.